

**REALISIERUNG VON WELLENDIGITALFILTERN
MIT GLEITKOMMAARITHMETIK**

DISSERTATION

zur

Erlangung des Grades eines

Doktor-Ingenieurs

der

Abteilung für Elektrotechnik

an der Ruhr-Universität Bochum

von

MOHAMED IBRAHIM YOUSSEF

Menofia (Ägypten)

Bochum 1988





Digitized by the Internet Archive
in 2026

**REALISIERUNG VON WELLENDIGITALFILTERN
MIT GLEITKOMMAARITHMETIK**

DISSERTATION

zur

Erlangung des Grades eines

Doktor-Ingenieurs

der

Abteilung für Elektrotechnik

an der Ruhr-Universität Bochum

von

MOHAMED IBRAHIM YOUSSEF

Menofia (Ägypten)

Bochum 1988

Dissertation eingereicht am: 1. 12.1987

Tag der mündlichen Prüfung: 22.1.1988

Referent: Prof. Dr. sc. techn. Dr. hc. A. Fettweis

Korreferent: Prof. Dr.-Ing. K. Meerkötter

Inhaltsverzeichnis

Seite

1. <u>Einleitung</u>	5
2. <u>Grundlagen der Wellendigitalfilter</u>	
2.1 Vorbemerkungen	9
2.2 Wellenflußdiagramm	11
2.3 Empfindlichkeit der Wellendigitalfilter	20
2.4 Abhängigkeit zwischen der Empfindlichkeit und dem Quantisierungsrauschen	30
2.5 Stabilitätsverhalten der Wellendigitalfilter	31
3. <u>Gleitkommazahlendarstellung</u>	
3.1 Vorbemerkungen	34
3.2 Wahrscheinlichkeitsdichtefunktion einer Gleitkommazahl	36
3.3 Quantisierungsfehler bei Gleitkommazahlen	38
3.4 Verteilungsdichtefunktion des relativen Quantisierungsfehlers	40
3.5 Wahrscheinlichkeitsdichtefunktion des absoluten Fehlers	45
3.6 Fehler durch Begrenzung des Zahlenbereiches	46
3.6.1 Exponentenüberlauf	46
3.6.2 Exponentenunterlauf	46
3.7 Zahlenbereich und die Genauigkeit von Gleit- und Festkommadarstellung	47
3.8 Die elementaren arithmetischen Operationen	50
3.8.1 Addition und Subtraktion von zwei Gleitkommazahlen	50
3.8.2 Multiplikation von zwei Gleitkommazahlen	51
3.9 Komplexität der Hardwarerealisierung einer Gleitkommaarithmetik	52
4. <u>Passivitätseigenschaften von Gleitkomma-Wellendigitalfiltern</u>	
4.1 Vorbemerkungen	56
4.2 Stabilitätskriterium	58
4.3 Pseudopassivitätseigenschaften der Bauelemente von WDF	61

4.3.1	Vorbemerkungen	61
4.3.2	Pseudopassivitätseigenschaften der Adaptoren	61
4.3.2.1	Pseudopassivität des 2-Tor-Adaptors	62
4.3.2.2	Pseudopassivität des 3-Tor-Adaptors	66
4.3.3	Pseudopassivitätseigenschaften des Übertragers	71
4.4	Inkrementale Pseudopassivität der Adaptoren und der Übertrager	75
4.5	Stabilitätseigenschaften von Bireziprok-Wellendigitalfilter	79
4.6	Filterbeispiele	82
5.	<u>Stabilitätsuntersuchungen von konventionellen rekursiven Digitalfiltern 2.Ordnung</u>	
5.1	Grenzyklussschwingungen bei abgeschal- tetem Eingangssignal	93
5.2	Ein auf Wellendigitalfilter beruhendes Gleitkomma- Digitalfilter 2.Ordnung ohne Grenzyklussschwingungen ..	100
5.3	Stabilität der Nachrichtenschleife	103
6.	<u>Berechnung des Signal-Rausch-Verhältnisses bei Gleitkomma- arithmetik</u>	
6.1	Vorbemerkungen	109
6.2	Berechnung des Rauschens eines Quantisierers	110
6.2.1	Rundungsrauschen	111
6.2.2	Rauschen bei Betragsschneiden	112
6.3	Vergleich zur Festkommaarithmetik	115
6.4	Berechnung des Rundungsrauschens eines beliebigen Digitalfilters	119
6.5	Beispiel	121
7.	<u>Zusammenfassung</u>	124
	<u>Literaturverzeichnis</u>	126

1. Einleitung

In den letzten Jahren hat die digitale Signalverarbeitung zunehmend an Bedeutung erlangt. Dies liegt im wesentlichen daran, daß Digitalerschaltungen wegen der heute erreichbaren großen Integrationsdichte bei der Herstellung integrierter Schaltungen kleiner und preiswerter als Analogschaltungen sind. Darüber hinaus weist die Digitalrealisierung eines Systems eine Anzahl von Vorteilen gegenüber der Analogrealisierung auf, wie z.B.:

- 1) keine Veränderungen der Systemeigenschaften durch Alterung und/oder Temperaturschwankungen,
- 2) das Toleranzproblem der Bauelemente tritt nicht auf,
- 3) hohe Zuverlässigkeit und geringe Störanfälligkeit,
- 4) hohe Flexibilität des Übertragungsverhaltens mit einer festen Anordnung.

In der digitalen Signalverarbeitung spielen Digitalfilter, die grundsätzlich aus numerisch operierenden Speichern, Multiplizierern und Addierern aufgebaut werden, eine wichtige Rolle. In einem Digitalfilter wird ein digitalisiertes Signal, das in der Regel durch Abtastung, Quantisierung und Kodierung aus einem zeitkontinuierlichen Signal entsteht, nach einem vorgeschriebenen Rechenalgorithmus verarbeitet und als Ausgangszahlenfolge ausgegeben. Es gibt prinzipiell zwei verschiedene Typen von Digitalfiltern, nämlich die rekursiven und die nichtrekursiven. Zur Implementierung des Rechenalgorithmus eines Digitalfilters kommen verschiedene Arithmetikarten in Betracht. Abgesehen von einigen Arithmetiktypen, die bislang noch keine praktische Bedeutung erlangten, haben Fest- und Gleitkommaarithmetik in der digitalen Signalverarbeitung ihre Anwendung gefunden.

Bekanntlich werden in einem Digitalfilter alle Filterkoeffizienten und Signalparameter mit einer begrenzten Wortlänge dargestellt. Hierdurch ergeben sich Fehler, die als Quantisierungsfehler bezeichnet werden und zu Abweichungen von dem gewünschten Systemverhalten führen. Diese Fehler können nach ihren Ursachen in zwei Kategorien unterteilt werden. Die erste Kategorie wird durch die notwendige Begrenzung der Koeffizientenwortlänge hervorgerufen (Einstellfehler) und führt zu einer Abweichung des Übertragungsverhaltens, z.B. des Dämpfungsverlaufs, von dem ursprünglich angestreb-

ten Verhalten. Die zweite Kategorie wird durch die Rundungs- und/oder Schneideooperationen nach bestimmten arithmetischen Operationen zur Begrenzung der Signalwortlänge (Nichtlinearitäten) verursacht. Diese lassen sich in Rundungsrauschen, parasitäre Schwingungen (Grenzzyklusschwingungen), Antwortinstabilitäten und Schleifeninstabilitäten einteilen [13]. Offensichtlich treten solche Schwingungen nur dann auf, wenn Rückkopplungen vorhanden sind, also ausschließlich bei rekursiven Digitalfiltern.

Neben den herkömmlichen Strukturen, die für Digitalfilter vorgeschlagen worden sind, haben die Wellendigitalfilter (WDF) eine größere Bedeutung erlangt [1]. Bekanntlich werden WDF-Strukturen aus klassischen, vorzugsweise verlustfreien und beidseitig resistiv abgeschlossenen Filtern, die als Referenzfilter bezeichnet werden, hergeleitet [1]. In diesem Fall übertragen sich einige gute Eigenschaften des Referenzfilters auf das dazugehörige WDF [25]. Als Folge der geringen Empfindlichkeit der Dämpfung gegenüber Bauelementvariationen im Referenzfilter können durch diskrete Optimierung die Koeffizienten eines WDF's für praktisch viele wichtige Fälle sehr einfach realisiert werden [16-18, 20]. Da ein direkter Zusammenhang zwischen der Dämpfungsempfindlichkeit eines Digitalfilters und dem an seinem Ausgang auftretenden Quantisierungsrauschen besteht [12,21], wirkt sich diese geringe Empfindlichkeit ferner günstig auf den Dynamikbereich eines Wellendigitalfilters aus [12,21,69].

Im Falle der Festkommaarithmetik lassen sich WDF auch unter nichtlinearen Bedingungen pseudopassiv realisieren. Mit dem Konzept der Pseudopassivität konnten technisch einfach zu realisierende Maßnahmen angegeben werden, die bewirken, daß WDF (im Gegensatz zu konventionellen rekursiven Digitalfiltern) bei abgeschaltetem Eingangssignal völlig frei von Grenzzyklusschwingungen realisierbar sind [25]. Dies kann sogar gewährleistet werden, wenn das Filter innerhalb einer Nachrichtenschleife arbeitet [50]. Es wurde ferner gezeigt, daß WDF unter Vernachlässigung der Signalgranularitäten inkremental pseudopassiv realisiert werden können [29,30]. Diese inkrementale Pseudopassivität führt zur Antwortstabilität und Stetigkeit des Übertragungsverhaltens eines WDF's unter nichtlinearen Bedingungen. Unter Antwortstabilität versteht man, daß die durch einen Überlauf verursachten Störungen am Filterausgang auch immer dann mit der Zeit abklingen müssen, wenn das Eingangssignal auf

einen beliebigen Pegel abgesenkt wird, bei dem unter linearen Bedingungen keine Exponentenüberläufe auftreten. Die Stetigkeit des Übertragungsverhaltens bedeutet, daß kleinere Änderungen des Eingangssignals und/oder der Anfangeswerte nur kleinere Änderungen des Ausgangssignals verursachen können, und im allgemeinen, daß geringfügige Übersteuerungen am Eingang nur geringfügige Verzerrungen am Ausgang des Filters zur Folge haben können. Durch diese interessanten Eigenschaften (nämlich die geringe Dämpfungsempfindlichkeit und die Pseudopassivität) sind die WDF den konventionellen rekursiven Digitalfiltern überlegen.

Die vorliegende Arbeit befaßt sich mit der Realisierung von WDF mit Gleitkommaarithmetik, d.h., daß sowohl Multipliziererkoeffizienten als auch Signalparameter als Gleitkommagrößen dargestellt werden. In Kapitel 2 dieser Arbeit werden zunächst einige für das Verständnis erforderliche Grundlagen aus der Theorie der WDF sowie einige günstige Eigenschaften dieser Digitalfilter kurz zusammengefaßt. Hieraus sind insbesondere die geringe Empfindlichkeit der Dämpfung gegenüber Koeffizientenänderungen und das Stabilitätsverhalten solcher Filter zu erwähnen.

In Kapitel 3 werden die Gleitkommazahldarstellung sowie die Quantisierungsfehler und der Zahlenbereich bei Gleitkommaarithmetik näher betrachtet. Außerdem werden in diesem Kapitel die elementaren arithmetischen Gleitkommaoperationen (zur Addition und Multiplikation zweier Gleitkommazahlen) und eine Schätzung der Komplexität der Hardwarerealisierung (im Falle eines Addierers und eines Multiplizierers bei Fest- und Gleitkommaarithmetik) dargestellt.

Ein großer Teil dieser Arbeit orientiert sich an den Pseudopassivitätseigenschaften von Wellendigitalfiltern, die mit Gleitkommaarithmetik realisiert werden. Wir werden uns auf den praktisch interessanteren Fall, nämlich den Fall der Darstellung einer Zahl als Gleitkommagröße in Zweierkomplementkode, beschränken. Es wird in Kapitel 4 gezeigt, daß mit Hilfe des Konzepts der Pseudopassivität technisch zu realisierende Maßnahmen angegeben werden, die bewirken, daß nach Abschalten des Eingangssignals alle auftretenden parasitären Schwingungen abklingen. Dies kann auch gewährleistet werden, wenn das Filter innerhalb einer Nachrichtenschleife arbeitet. Außerdem wird in diesem Kapitel mit dem Konzept der inkrementalen Pseudopassivität nachgewiesen, daß bei Verwendung geeigneter Exponentenüberlaufcharakteristiken (unter Vernachlässigung der

Signalgranularität) Gleitkomma-Wellendigitalfilter auch antwortstabil sind. Es zeigt sich also, daß sich bei solchen Filtern die Stabilitätsprobleme vollständig lösen lassen. Weiterhin werden wir die Stabilitätseigenschaften von Bireziprok-Wellendigitalfiltern bei abgeschaltetem Eingangssignal diskutieren. Es hat sich herausgestellt, daß solche Filterstrukturen und Polyphasenfilter, die unter Verwendung von Wellendigitalfiltern realisiert werden [60], durch sehr einfache Maßnahmen bei abgeschaltetem Eingangssignal völlig frei von Grenzzyklusschwingungen realisiert werden.

Da sich Digitalfilter konventioneller Art als Kaskadenschaltung von Gliedern 1. und 2. Ordnung realisieren lassen, um mit einer kürzeren Wortlänge für die Multipliziererkoeffizienten auskommen zu können, befaßt sich ein weiterer Teil der vorliegende Arbeit in Kapitel 5 mit den Stabilitätsuntersuchungen von konventionellen rekursiven Digitalfiltern zweiter Ordnung. Diese Untersuchungen orientieren sich im wesentlichen an der Stabilität sowohl bei abgeschaltetem Eingangssignal als auch unter Schleifenbedingungen. Es wird sich zeigen, daß Grenzzyklusschwingungen mit teilweise großen Amplituden in solchen Filtern unter den oben genannten Betriebsbedingungen entstehen können. Außerdem wird in diesem Kapitel ein konventionelles Gleitkomma-Digitalfilter 2. Ordnung ohne Grenzzyklusschwingungen bei abgeschaltetem Eingangssignal und ohne Antwortinstabilitäten besprochen. Ein solches Glied ist noch stets kanonisch bezüglich der Anzahl der Verzögerungsglieder. Die Anzahl der Addierer und der Multiplizierer wird jeweils von 4 auf 6 erhöht.

In Kapitel 6 wird die Untersuchung des Rauschens von Digitalfiltern mit Hilfe des Modells unkorrelierter Rauschquellen untersucht. Es wird sich herausstellen, daß das Signal-Rausch-Verhältnis am Ausgang eines Digitalfilters bei Gleitkommaarithmetik ein vorteilhaftes Verhalten gegenüber Festkommaarithmetik besitzt, wenn der Aufwand der Hardwarerealisierung in Kauf genommen wird.

2. Grundlagen der Wellendigitalfilter

2.1 Vorbemerkungen

Digitalfilterstrukturen, die auf der Nachbildung von klassischen Filternetzwerken beruhen, nennt man Wellendigitalfilter. Das klassische Filter, von dem man das entsprechende Wellendigitalfilter herleitet, wird als Referenzfilter bezeichnet. Die von Fettweis geschriebene Arbeit [1] gibt eine ausführliche Übersicht der Wellendigitalfilter und beinhaltet die wichtigsten Grundlagen und Formeln für solche Digitalfilter.

In der Theorie der Wellendigitalfilter werden nicht die üblichen physikalischen Strom- und Spannungsgrößen, die im klassischen Filter auftreten, als Signalgrößen verwendet, sondern die sogenannten Wellengrößen. Unter Wellen versteht man allgemeine bestimmte Linearkombinationen von Strom-, Spannungs- und Leistungsgrößen. Die Verwendung der im Referenzfilter auftretenden Spannungen und Ströme führt im allgemeinen zu einem nicht realisierbaren Signalflußdiagramm, welches verzögerungsfreie Schleifen enthält [2,7].

Die Wahl zwischen Strom- und Spannungswellen ist irrelevant, da ein Wellendigitalfilter, das von einem Referenzfilter unter Verwendung von Spannungswellen nachgebildet wird, identisch ist mit einem Wellendigitalfilter, das von dem entsprechenden dualen Referenzfilter unter Verwendung von Stromwellen hergeleitet wird [5]. In dieser Arbeit beschränken wir uns auf den Fall der Spannungswellen als Signalgrößen. Bild 2.1 stellt ein beliebig gewähltes Tor U dar, dem ein Torwiderstand R_U in Form einer beliebig gewählten positiven Konstanten zugeordnet wird. Weiterhin sind im Bild 2.1 die Wellengrößen a_U als hinlaufende und b_U als rücklaufende Welle eingezeichnet.

Die Wellengrößen a_U und b_U werden dann durch die folgenden Beziehungen [2]

$$a_U = v_U + i_U R_U \quad , \quad (2.1a)$$

$$b_U = v_U - i_U R_U \quad , \quad (2.1b)$$

definiert, wobei v_U der Momentanwert der Spannung und i_U der Momentanwert des Stroms ist. Die komplexen Amplituden des stationären Zustands bezeichnet man mit den entsprechenden großen Buchstaben.

Statt der üblichen komplexen Frequenz $P = \sigma + j\omega$ wird die Frequenzvariable ψ im Referenzfilter verwendet. Dieses ψ spielt in der Theorie der Digitalfilter, insbesondere der Wellendigitalfilter,

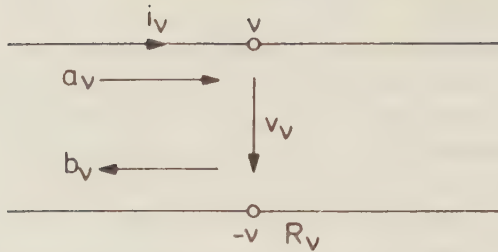


Bild 2.1: Ein Tor $\cup$ mit den Klemmen $\cup$ und $-\cup$, dem Torwiderstand $R_{\cup}$, der Spannung $v_{\cup}$, dem Strom $i_{\cup}$, der hinlaufenden Welle $a_{\cup}$ und der rücklaufenden Welle $b_{\cup}$.

eine gewisse Rolle, wenn es mit der üblichen komplexen Frequenzvariablen p und mit der Betriebsperiode T durch die folgende Beziehung verbunden wird:

$$\psi = \tanh(pT/2) = \frac{e^{pT}-1}{e^{pT}+1} = \frac{z-1}{z+1} \quad , \quad (2.2)$$

dabei ist $z=e^{pT}$ die aus der z -Transformation bekannte Größe. Für eine reelle Frequenz, d.h. $p=j\omega$, ergibt sich der Zusammenhang zwischen der physikalischen Frequenz ω und der entsprechenden Frequenzvariablen ψ in dem Referenzfilter als

$$\psi = j\phi = \tanh\left(\frac{j\omega T}{2}\right) = j\tan\left(\frac{\omega T}{2}\right) \quad (2.3a)$$

oder

$$\phi = \tan\left(\frac{\omega T}{2}\right) \quad (2.3b)$$

Der Gleichung (2.3b) entnehme man, daß jeweils Abschnitte der Länge $\frac{2\pi}{T}$ auf der ω -Achse auf die gesamte ϕ -Achse abgebildet werden.

Das bedeutet, daß der Grundabschnitt auf der ω -Achse (Nyquist range $0 < |\omega| < \pi/T$) exakt auf die ϕ -Achse ($0 < |\phi| < \infty$) abgebildet wird. Die Betrachtung des Grundabschnitts genügt, da das Spektrum sich auf der ω -Achse im Abstand von $\frac{2\pi}{T}$ wiederholt.

Darüber hinaus gewinnt man von der definierten Transformation (2.2) die folgenden Eigenschaften [1]:

$$\begin{array}{lll} \operatorname{Re} \psi > 0 & \operatorname{Re} p > 0 & |z| > 1, \\ \operatorname{Re} \psi < 0 & \operatorname{Re} p < 0 & |z| < 1. \end{array} \quad (2.4)$$

Hieraus wird ersichtlich, daß ein stabiles kausales Digitalfilter nach einem stabilen Referenzfilter und umgekehrt aufgebaut wird.

2.2 Wellenflußdiagramm

Der Begriff des Wellenflußdiagramms wird anstatt des Signalflußdiagramms benutzt, da Wellen als Signalgrößen verwendet werden. Um ein Wellenflußdiagramm für ein Wellendigitalfilter herleiten zu können, zerlegt man das nachzubildende Referenzfilter in seine Bauelemente sowie in die verbleibende Verbindungsstruktur. Für jedes im Referenzfilter auftretende Bauelement wird dann das entsprechende Wellendigitalelement angegeben. Anhand eines Beispiels kann man diese Umwandlung der Bauelemente vom Referenzfilter ins Wellendigitalfilter erläutern. Nun betrachten wir im Referenzfilter eine Kapazität mit der Impedanz $Z=R/\psi$ (Bild 2.2a). Die Spannung V wird beschrieben durch

$$V = \frac{R}{\psi} I \quad (2.5)$$

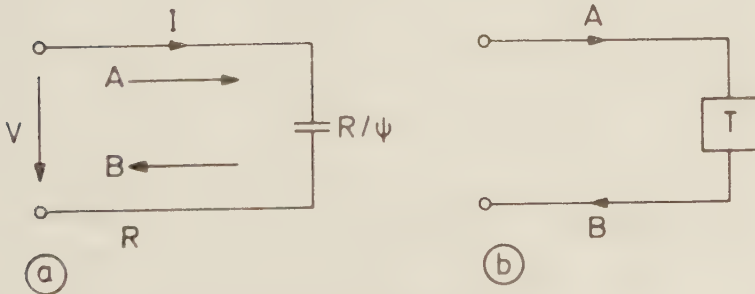


Bild 2.2: a) Eine Kapazität im Referenzfilter.
b) Das zugehörige Wellenflußdiagramm.

Um eine möglichst einfache Digitalrealisierung zu erhalten, wird der Torwiderstand gleich dem R gewählt. In diesem Fall erhalte man durch die Definitionsgleichungen (2.1) unter Verwendung von (2.5) die folgende Beziehung

$$B = z^{-1} A$$

(2.6)

Wie man aus der Beziehung (2.6) ersehen kann, wird im Wellendigitalfilter aus dieser Kapazität ein Verzögerer (Bild 2.2b).

In der Tabelle 2.1 sind einige der reaktiven Bauelemente, nämlich Kapazität, Induktivität und QUARL (nach dem englischen quasi reciprocal line) mit ihren Definitionsgleichungen und den daraus folgenden Wellenflußdiagrammen dargestellt [1]. Die Tabelle 2.2 zeigt die wichtigsten nichtreaktiven Bauelemente, und die Tabelle 2.3 enthält die Wellenflußdiagramme der Quellen.

Für die Verbindungsstruktur werden mittels der Kirchhoffschen Gesetze und der Beziehungen (2.1) die sogenannten Adaptoren angegeben. Über diese Adaptoren werden die Wellendigitalelemente zusammengeschaltet.

Element	Schaltbild	Wellen - beziehung	Wellenfluß - diagramm
Kapazität		$b_1(t_m) = a_1(t_m - T)$	
Induktivität		$b_1(t_m) = -a_1(t_m - T)$	
Leitungselement (QUARL)		$b_1(t_m) = a_2(t_m)$ $b_2(t_m) = a_1(t_m - T)$	

Tabelle 2.1: Wellenflußdiagramme einiger reaktiven Bauelemente mit ihren Wellenbeziehungen ($t_m = t_o + mT$, $m = \dots -1, 0, 1, 2, \dots$ und t_o ist beliebige Konstante).

Element	Schaltbild	Wellenbeziehung	Wellenflußdiagramm
Widerstand		$b_1(t_m) = 0$	
Kurzschluß		$b_1(t_m) = -a_1(t_m)$ (R beliebig)	
Leerlauf		$b_1(t_m) = a_1(t_m)$ (R beliebig)	
Gyrator		$b_1(t_m) = -a_2(t_m)$ $b_2(t_m) = a_1(t_m)$	
3-Tor-Zirkulator		$b_1(t_m) = a_3(t_m)$ $b_2(t_m) = a_1(t_m)$ $b_3(t_m) = a_2(t_m)$	

Tabelle 2.2: Wellenflußdiagramme der wichtigsten nichtreaktiven Bauelemente mit ihren Wellenbeziehungen.

Neben der Realisierung eines bestimmten Verbindungstyps haben die Adaptoren die Aufgabe, die unterschiedlichen Torwiderstände der Bauelemente anzupassen. In diesen Adaptoren werden alle benötigten arithmetischen Operationen, nämlich die Addition und die Multiplikation von Signalen mit konstanten Koeffizienten ausgeführt. Infolge der endlichen Koeffizienten- und Signalwortlänge treten bei der Berechnung der Ausgänge eines Adaptors Fehler auf. Deshalb müssen wir diesen Adaptoren besondere Aufmerksamkeit widmen, um ein stabiles, unter allen Betriebsbedingungen arbeitendes Wellendigitalfilter zu realisieren.

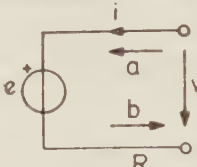
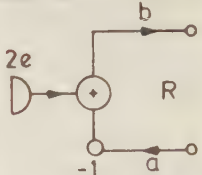
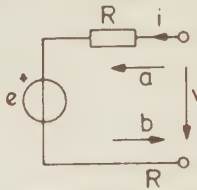
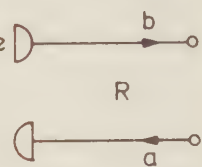
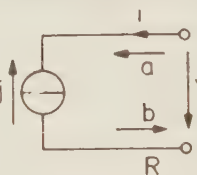
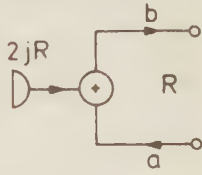
Quelle	Schaltbild	Wellenbe- ziehung	Wellenfluß- diagramm
Ideale Spannungs- quelle		$b = 2e - a$ (R beliebig)	
Spannungs- quelle mit Widerstand		$b = e$	
Ideale Stromquelle		$b = 2jR + a$ (R beliebig)	

Tabelle 2.3 : Wellenflußdiagramme einiger Quellen.

Bei der Zusammenschaltung von Adaptoren ist darauf zu achten, daß die beiden zu verbindenden Tore den gleichen Torwiderstand und die gleiche Wellenrichtung besitzen. Außerdem dürfen keine verzögerungsfreien Schleifen auftreten. Dies wird entweder nach [2], durch Schalten eines QUARL's zwischen die zwei zu verbindenden Tore oder nach [6], durch die "Reflexionsfreiheit" eines Tores der zwei zu verbindenden Tore, erreicht.

Die Ermittlung der Adaptorschaltungen wird folgendermaßen erläutert. Die Parallelverbindung von n Toren, die im Referenzfilter auftritt, wird durch einen n-Tor-Paralleladaptor im Wellendigitalfilter realisiert (Bild 2.3). Die im Referenzfilter auftretende Reihenverbindung von n-Toren wird durch einen n-Tor-Reihenadaptor realisiert (Bild 2.4). Die zugehörige n-Tor-Paralleladaptor-Gleichung [1] läßt sich mit der Gültigkeit der Kirchhoffschen Gesetze der Par-

allelverbindung

$$v_1 = v_2 = \dots = v_n \text{ und } i_1 + i_2 + \dots + i_n = 0$$

sowie mit den Definitionsgleichungen (2.1) berechnen durch

$$b_u = \sum_{i=1}^n \gamma_i a_i - a_u, \quad u = 1, 2, \dots, n \quad (2.7)$$

Die Koeffizienten γ_i in (2.7) werden bestimmt aus

$$\gamma_i = \frac{2G_i}{G}, \quad G_i = \frac{1}{R_i} \quad (2.8)$$

mit

$$G = \sum_{i=1}^n G_i, \quad (2.9)$$

wobei der Torwiderstand R_i dem Tor i zugeordnet ist. Aus (2.8) und (2.9) ergibt sich die Bedingung

$$\sum_{i=1}^n \gamma_i = 2 \quad (2.10)$$

Auf ähnliche Weise ergeben sich die n -Tor-Reihenadapter-Gleichungen [1]

$$b_u = a_u - \gamma_u \sum_{i=1}^n a_i, \quad u = 1, 2, \dots, n \quad (2.11)$$

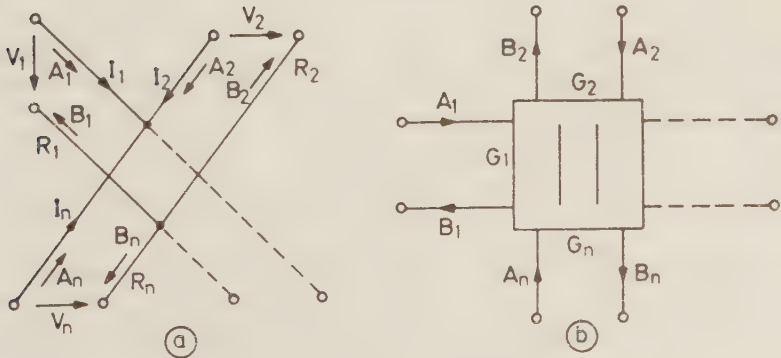


Bild 2.3: a) Parallelverbindung von n -Toren,
b) n -Tor-Paralleladapter.

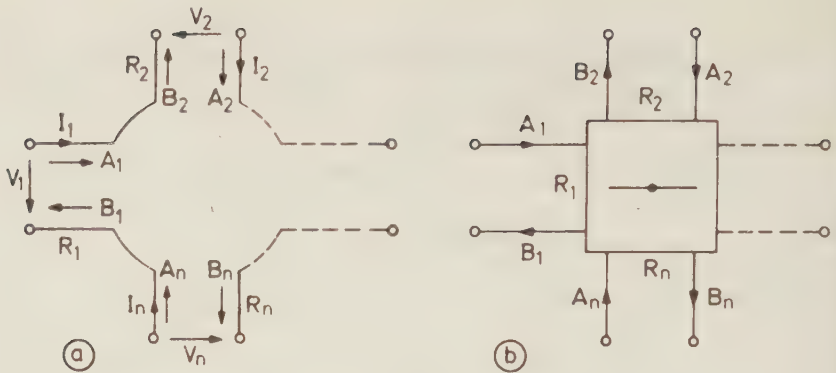


Bild 2.4: a) Reihenverbindung von n-Toren,
 b) n-Tor-Reihenadaptor.

Die Koeffizienten γ_U berechnen sich in diesem Fall aus

$$\gamma_U = \frac{2R_U}{R} \quad \text{mit} \quad R = \sum_{U=1}^n R_U \quad (2.12)$$

Die Bedingung (2.10) gilt auch für die Koeffizienten im Fall des Reihenadaptors.

Die Reflexionsfreiheit eines Tores kann bei einem Adaptor durch eine geeignete Wahl des betreffenden Torwiderstandes erzielt werden [6]. Im Fall des Reihenadaptors muß der Widerstand des reflexionsfreien Tores gleich der Summe der übrigen Widerstände

sein. Sei Tor n reflexionsfrei, d.h., $R_n = \sum_{U=1}^{n-1} R_U$, so erhalten wir

aus (2.12) die Gleichungen

$$\gamma_n = 1, \quad \gamma_U = \frac{R_U}{R_n}, \quad U = 1, 2, \dots, n-1 \quad (2.13)$$

Die entsprechende Aussage für Torleitwerte gilt für den Paralleladaptor.

Für den praktischen Wellendigitalfilterentwurf sind die 3-Tor- und die 2-Tor-Adaptoren wichtig (vgl. Bild 2.5). Der 2-Tor-Adaptor wird benötigt, wenn zwei Tore, die unterschiedliche Torwiderstände aufweisen, miteinander verbunden werden sollen. Darüber hinaus spie-

len sie bei der Realisierung von Brücken-Wellendigitalfiltern als Kaskadenschaltungen von Allpässen erster und zweiter Ordnung oder Kettenschaltungen von Leitungselementen die Hauptrolle. Dieser 2-Tor-Adaptor kann sowohl als 2-Tor-Paralleladaptor nach (2.7) als auch als 2-Tor-Reihenadaptor nach (2.11) realisiert werden.

Für einen allgemeinen 3-Tor-Paralleladaptor läßt sich (2.7) folgendes schreiben:

$$\begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} = \begin{bmatrix} \gamma_1 - 1 & \gamma_2 & \gamma_3 \\ \gamma_1 & \gamma_2 - 1 & \gamma_3 \\ \gamma_1 & \gamma_2 & \gamma_3 - 1 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} \quad . \quad (2.14)$$

Auf gleicher Weise erhält man für einen 2-Tor-Paralleladaptor

$$\begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} -\gamma & \gamma_2 \\ \gamma_1 & \gamma \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} \quad , \quad (2.15)$$

mit

$$\gamma = \frac{R_1 - R_2}{R_1 + R_2} = 1 - \gamma_1 = \gamma_2 - 1 \quad .$$

Für einen allgemeinen 3-Tor-Reihenadaptor ergibt sich die folgende Beziehung

$$\begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} = \begin{bmatrix} 1 - \gamma_1 & -\gamma_1 & -\gamma_1 \\ -\gamma_2 & 1 - \gamma_2 & -\gamma_2 \\ -\gamma_3 & -\gamma_3 & 1 - \gamma_3 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} \quad . \quad (2.16)$$

Entsprechend gilt für einen 2-Tor-Reihenadaptor

$$\begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} -\gamma & -\gamma_1 \\ -\gamma_2 & \gamma \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} \quad , \quad (2.17)$$

mit

$$\gamma = \frac{R_1 - R_2}{R_1 + R_2} = \gamma_1 - 1 = 1 - \gamma_2 \quad .$$

Im Kapitel 4 werden wir sehen, daß die Realisierung eines Adaptors direkt durch seine Streumatrix im Fall der Gleitkommaarithmetik günstig ist, um die Pseudopassivitätseigenschaften einfach si-

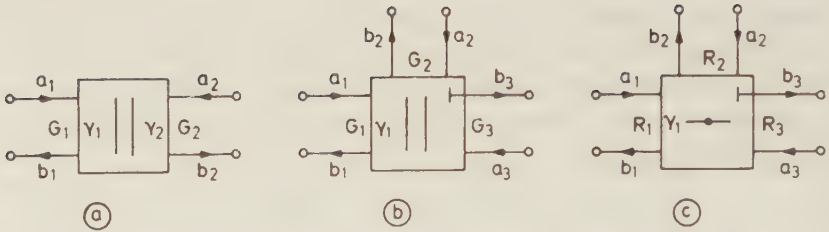


Bild 2.5: a) 2-Tor-Adaptor, b) 3-Tor-Paralleladaptor und c) 3-Tor-Reihenadaptor.

cherzustellen.

Ein weiteres Wellendigitalbauelement ist noch der ideale Übertrager, dessen Realisierungsmöglichkeiten in der Tabelle 2.4 gezeigt werden.

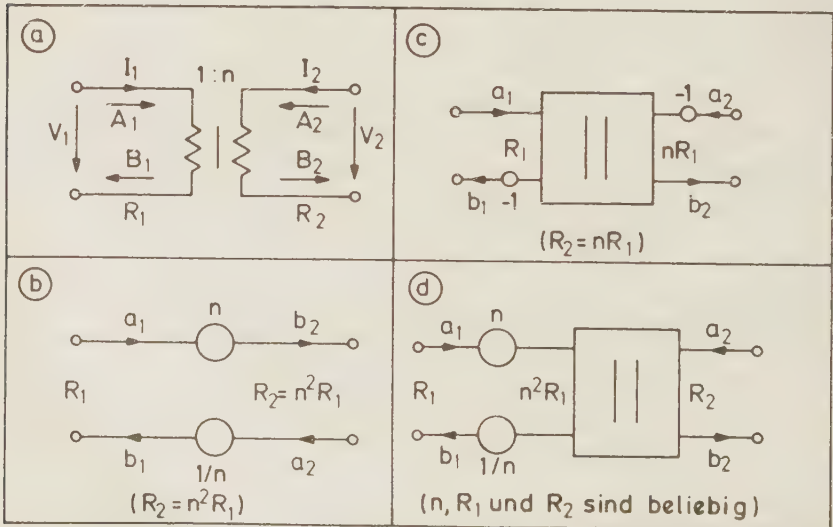


Tabelle 2.4: a) Ein idealer Übertrager, b), c) und d) drei Realisierungsmöglichkeiten [1].

Anhand eines Beispiels ist in Bild 2.6 der Prozeß der Herleitung eines Wellendigitalfilters von einem Tiefpaß-Referenzfilter in Abzweigstruktur ausführlich erläutert. Das Verbindungsnetzwerk des

Referenzfilters (Bild 2.6a) ist in Bild 2.6b dargestellt. Diese Verbindungen werden nun mit Hilfe von 3-Tor-Parallel- und Reihendapto-

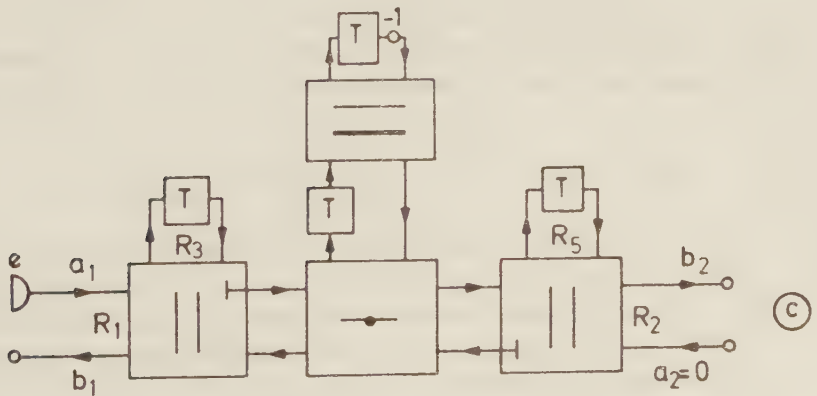
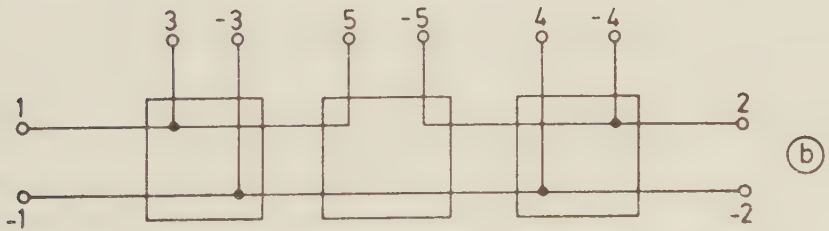
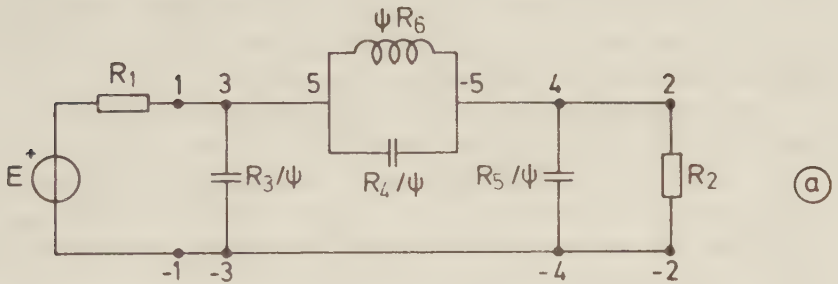


Bild 2.6: Beispiel für die Herleitung eines Wellendigitalfilters
 a) Tiefpaß-Referenzfilter,
 b) zugehöriges Verbindungsnetzwerk,
 c) vollständiges Wellendigitalfilter.

ren unter Berücksichtigung der Reflexionsfreiheit eines von zwei direkt miteinander verbundenen Toren nachgebildet. Wenn wir die Tore dieser Anordnung noch mit den entsprechenden Elementen und der Quelle zusammenschalten, dann erhalten wir das in Bild 2.6c vollständig dargestellte Wellendigitalfilter.

Man kann sich leicht überzeugen, daß ein Wellendigitalfilter, das aus einem verlustfreien Referenzfilter hergeleitet ist, von Natur aus ein Weichenfilter ist, da zwei Ausgänge, die im Bild 2.6c mit b_1 und b_2 eingezeichnet sind, als Übertragungsfunktionen zur Verfügung stehen und diese über die Feldtkeller-Gleichung miteinander verknüpft sind.

2.3 Empfindlichkeit der Wellendigitalfilter

Reaktanzfilter, die beidseitig resistiv abgeschlossen sind, weisen sehr günstige Empfindlichkeitseigenschaften der Dämpfung gegenüber Bauelementenvariationen auf [8]. Da Wellendigitalfilter nach den klassischen Filternetzwerken, besonders nach den von beidseitig resistiv abgeschlossenen Abzweig-Referenzfiltern, aufgebaut werden und zwischen den Multipliziererkoeffizienten eines Wellendigitalfilters und den Elementen des zugehörigen Referenzfilters ein umkehrbar eindeutiger Zusammenhang besteht, läßt sich die geringere Empfindlichkeit der Dämpfung gegenüber Multipliziererkoeffizientenvariationen in einem Wellendigitalfilter beweisen. Es empfiehlt sich für jeden Multipliziererkoeffizienten

γ_i die Empfindlichkeit $\frac{\partial \alpha}{\partial \gamma_i}$ zu untersuchen, wobei α die

Betriebsdämpfung eines gewünschten Toleranzschemas ist.

Nun betrachten wir das zum Wellendigitalfilter zugehörige beidseitig resistiv abgeschlossene Reaktanz-Referenzfilter (Bild 2.7). Die Dämpfungsänderung bezüglich der Variation eines Elements läßt sich nun folgendermaßen abschätzen.

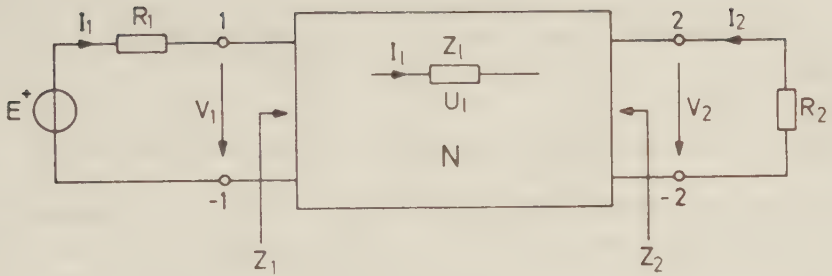


Bild 2.7: Ein verlustfreies Zweitorfilter.

Die Transmittanz S_{21} lässt sich ausdrücken durch

$$S_{21} = 2 \sqrt{\frac{R_1}{R_2}} \frac{V_2}{E} = e^{-\Gamma} = C S'_{21} \quad (2.18a)$$

mit

$$\Gamma = \alpha + j\beta, \quad C = \sqrt{\frac{R_1}{R_2}} \quad \text{und} \quad S'_{21} = \frac{2V_2}{E}. \quad (2.18b)$$

Die Reflektanzen lassen sich berechnen durch

$$S_{11} = \frac{Z_1 - R_1}{Z_1 + R_1}, \quad S_{22} = \frac{Z_2 - R_2}{Z_2 + R_2}. \quad (2.19)$$

Man nehme an, daß ein Element von seinem nominellen Wert abweicht, z.B. Z_ℓ wird $Z_\ell + \Delta Z_\ell$. Demzufolge wird Γ in $\Gamma + \Delta\Gamma$ übergehen. Dann ist

$$\Delta\Gamma = \Delta\alpha + j\Delta\beta \quad (2.20)$$

Aus (2.18b) und (2.18a) habe man

$$\frac{\partial \Gamma}{\partial Z_\ell} = \frac{\partial \alpha}{\partial Z_\ell} + j \frac{\partial \beta}{\partial Z_\ell} = \frac{1}{E} \frac{\partial E}{\partial Z_\ell}, \quad (2.21)$$

unter der Annahme, daß die Ausgangsspannung V_2 konstant bleibt.

Nach Bild 2.7 erhalten wir

$$E = V_1 + I_1 R_1,$$

dann läßt sich $\frac{\partial \Gamma}{\partial Z_\ell}$ schreiben durch

$$\frac{\partial \Gamma}{\partial Z_\ell} = \frac{1}{V_1 + I_1 R_1} \left(\frac{\partial V_1}{\partial Z_\ell} + R_1 \frac{\partial I_1}{\partial Z_\ell} \right) \quad (2.22)$$

Aufgrund der Gültigkeit der Kirchhoffschen Regeln gilt die Beziehung

$$\sum_{i=1}^K U_i I_i = 0 \quad (2.23)$$

wobei K die Nummer der Zweige des zur in Bild 2.7 dargestellten Schaltung gehörenden Graphs bezeichnet und $U_1 = -V_1$ sowie $U_2 = -V_2$ ist. Dieser Satz beinhaltet das Gesetz über die Erhaltung der Energie [11].

Im Fall, wenn ein Element z.B. Z_ℓ von seinem nominellen Wert abweicht, ändern sich die Spannungen und die Ströme, nämlich U_i wird $U_i + \Delta U_i$ und I_i wird $I_i + \Delta I_i$. Der Satz von Tellegen kann dann folgendermaßen formuliert werden:

$$\sum_{i=1}^K U_i (I_i + \Delta I_i) = 0 \quad \text{und} \quad \sum_{i=1}^K I_i (U_i + \Delta U_i) = 0 \quad (2.24)$$

Der Beweis steht in [11]. Unter der Verwendung von (2.23) läßt sich (2.24) schreiben als

$$\sum_{i=1}^K U_i \Delta I_i = 0 \quad \text{und} \quad \sum_{i=1}^K I_i \Delta U_i = 0$$

oder im allgemeinen als

$$\sum_{i=1}^K (U_i \Delta I_i - I_i \Delta U_i) = 0 \quad (2.25)$$

Nimmt man an, daß die Spannung E so eingestellt wurde, daß V_2 und I_2 Konstante bleiben, so erhalten wir

$$V_1 \Delta I_1 - I_1 \Delta V_1 = \sum_{i=3}^K (U_i \Delta I_i - I_i \Delta U_i) \quad (2.26)$$

Da die konjugiert komplexen Größen I_i^* von I_i und U_i^* von U_i ebenfalls die Kirchhoffschen Gleichungen erfüllen, läßt sich (2.23) schreiben als

$$\sum_{i=1}^K U_i I_i^* = 0 \quad , \quad \sum_{i=1}^K U_i^* I_i = 0 \quad (2.27)$$

Diese Formulierung ist gleichbedeutend mit dem Gesetz über die Erhaltung der komplexen Leistung. Die folgende Beziehung kann auf ähnliche Weise wie (2.25) mit Hilfe von (2.27) bewiesen werden.

$$\sum_{i=1}^K (U_i^* \Delta I_i + I_i^* \Delta U_i) = 0 \quad . \quad (2.28)$$

Unter derselben Annahme, daß V_2 und I_2 Konstante bleiben, erhalten wir

$$V_1^* \Delta I_1 + I_1^* \Delta V_1 = \sum_{i=3}^K (U_i^* \Delta I_i + I_i^* \Delta U_i) \quad . \quad (2.29)$$

Besteht das N-Tor aus verlustfreien reziproken Elementen, d.h., aus Induktivitäten, Kapazitäten und idealen Übertragern, so können die beiden Gleichungen (2.26) und (2.29) übergehen in die Formen

$$V_1 \Delta I_1 - I_1 \Delta V_1 = - \sum_{i=3}^K I_i^2 \Delta Z_i \quad , \quad (2.30a)$$

$$V_1^* \Delta I_1 + I_1^* \Delta V_1 = \sum_{i=3}^K |I_i|^2 \Delta Z_i \quad , \quad (2.30b)$$

dabei ist angenommen worden, daß die differenziellen Terme zweiter Ordnung vernachlässigbar sind und keine Änderungen in idealen Übertragern auftreten.

Betrachten wir nun den Fall, wenn nur ein Element Z_ℓ von seinem Nominalwert mit ΔZ_ℓ abweicht, so erhalten wir aus (2.30)

$$I_1 \Delta V_1 - V_1 \Delta I_1 = I_\ell^2 \Delta Z_\ell \quad (2.31a)$$

$$I_1^* \Delta V_1 + V_1^* \Delta I_1 = |I_\ell|^2 \Delta Z_\ell \quad . \quad (2.31b)$$

Lösen wir die beiden Beziehungen (2.31a) und (2.31b) nach

$\frac{\Delta V_1}{\Delta Z_\ell}$ und $\frac{\Delta I_1}{\Delta Z_\ell}$, so bekommen wir

$$\frac{\partial V_1}{\partial Z_\ell} = \lim_{\Delta Z_\ell \rightarrow 0} \frac{\Delta V_1}{\Delta Z_\ell} = \frac{I_\ell (V_1 I_\ell^* + V_1^* I_\ell)}{2P_1} \quad , \quad (2.32a)$$

$$\frac{\partial I_1}{\partial Z_\ell} = \lim_{\Delta Z_\ell \rightarrow 0} \frac{\Delta I_1}{\Delta Z_\ell} = \frac{I_\ell (I_1^* I_\ell - I_1 I_\ell^*)}{2P_1} \quad , \quad (2.32b)$$

wobei P_1 die am Eingang des Zweitordfilters, N , gegebene Wirkleistung bezeichnet.

Setzen wir (2.32) in (2.17) ein, so läßt sich $\frac{\partial \Gamma}{\partial Z_\ell}$ wie folgt ausdrücken:

$$\frac{\partial \Gamma}{\partial Z_\ell} = \frac{I_\ell}{2P_1} (I_\ell^* + \frac{V_1^* - I_1^* R_1}{V_1 + I_1 R_1} I_\ell) \quad (2.33)$$

Unter Verwendung von (2.19) entnehme man aus (2.33) die folgende Beziehung

$$\frac{\partial \Gamma}{\partial Z_\ell} = \frac{1}{2P_1} (|I_\ell|^2 + \frac{E^*}{E} S_{11}^* I_\ell^2) \quad (2.34)$$

Die Empfindlichkeit $\frac{\partial \alpha}{\partial X_\ell}$ läßt sich folgendermaßen schreiben:

$$\frac{\partial \alpha}{\partial X_\ell} = \frac{1}{2P_1} |S_{11}| |I_\ell|^2 \sin \theta \quad (2.35)$$

wobei θ die Phase von $\frac{E^*}{E} S_{11}^* I_\ell^2$ ist.

Da zwischen den Multipliziererkoeffizienten eines Wellendigitalfilters und den Elementen des zugehörigen Referenzfilters ein umkehrbar eindeutiger Zusammenhang besteht, können wir aus der Beziehung (2.35) entnehmen, daß die Änderung der Betriebsdämpfung im Durchlaßbereich gegenüber Variationen der Filterkoeffizienten klein ist. Für praktische Fälle gilt im Durchlaßbereich $|S_{11}| \ll 1$, so daß auch außerhalb der Dämpfungsnullstellen ($S_{11}=0$) die Empfindlichkeit sehr gering ist.

Für ein Wellendigitalfilter ist die Transmittanz S'_{21} und die Betriebsdämpfung

$$\alpha' = \alpha + \ln C \quad \text{mit} \quad \alpha' = -\ln |S'_{21}| \quad (2.36)$$

von praktischem Interesse [12]. Diese unterscheidet sich von α nur durch die frequenzunabhängige Größe $\ln C$. Zum vollen Verständnis des mit (2.36) diskutierten Mechanismus ist wichtig hinzuweisen, daß eine Vertikalverschiebung eines Dämpfungsverlaufs um eine frequenzunabhängige Größe $\ln C$ keine Dämpfungsverzerrung verursacht [13].

Betrachten wir nun den schlechtesten Fall, wenn alle Elemente von ihren Nominalwerten abweichen, so läßt sich die Dämpfungsänderung bezüglich dieser Elemente folgendermaßen abschätzen. Ausgehend von der Beziehung (2.30) und unter der Annahme, daß alle reaktiven

Elemente des Referenzfilters die gleichen Toleranzen $\Delta Z_{in} = \frac{\Delta Z_i}{Z_i}$

aufweisen, erhalten wir

$$I_1 \Delta V_1 - V_1 \Delta I_1 = \sum_{i=3}^K I_i^2 Z_i (\Delta Z_{in}) \quad , \quad (2.37a)$$

und

$$I_1^* \Delta V_1 + V_1^* \Delta I_1 = \sum_{i=3}^K I_i I_i^* Z_i (\Delta Z_{in}) \quad . \quad (2.37b)$$

Lösen wir (2.37) nach $\Delta V_1 / \Delta Z_{in}$ und $\Delta I_1 / \Delta Z_{in}$, so bekommen wir

$$\frac{\partial V_1}{\partial Z_{in}} = \lim_{\Delta Z_{in} \rightarrow 0} \frac{\Delta V_1}{\Delta Z_{in}} = \frac{\sum_{i=3}^K (V_1^* I_i^2 + V_1 I_i I_i^*) Z_i}{2P_1} \quad (2.38a)$$

$$\frac{\partial I_1}{\partial Z_{in}} = \lim_{\Delta Z_{in} \rightarrow 0} \frac{\Delta I_1}{\Delta Z_{in}} = \frac{\sum_{i=3}^K (I_1 I_i I_i^* - I_1^* I_i^2) Z_i}{2P_1} \quad . \quad (2.38b)$$

Setzen wir (2.38) in (2.17) ein, so erhalten wir

$$\frac{\partial \Gamma}{\partial Z_{in}} = \frac{\sum_{i=3}^K (|I_i|^2 + \frac{E^*}{E} S_{11}^* I_i^2) Z_i}{2P_1} \quad . \quad (2.39)$$

Der Gleichung (2.39) kann man entnehmen, daß die folgende Beziehung gilt:

$$\Delta \Gamma = \Delta \alpha + j \Delta \beta = \sum_{i=3}^K \frac{(|I_i|^2 + \frac{E^*}{E} S_{11}^* I_i^2) \Delta Z_i}{2P_1}$$

und daraus folgt

$$\Delta \alpha = \frac{1}{2P_1} |S_{11}| \sum_{i=3}^K |I_i|^2 |\Delta Z_i| \sin \theta_i \quad , \quad (2.40)$$

wobei θ_i die Phase zwischen $\frac{E}{E} S_{11}$ und $I_i^2 \Delta Z_i$ bezeichnet.

Es ist einfach aus (2.40) zu ersehen, daß im allgemeinen die Empfindlichkeit der Betriebsdämpfung eines Wellendigitalfilters sehr gering ist, solange die Reflektanz S_{11} sehr klein ist.

Eine Methode, um die Empfindlichkeit eines Wellendigitalfilters gegenüber Variationen der Multipliziererkoeffizienten zu berechnen, ist in [14,15] gegeben.

Im Sperrbereich lassen sich die Dämpfungspole auf zwei grundsätzlich verschiedene Arten erzeugen. Im Fall eines Wellendigitalfilters in Abzweigstruktur werden sie durch die Resonanzfrequenz der Schwingkreise in den Serien- oder Parallelzweigen ermittelt. Bei solchen Filtern wird jeder Dämpfungspol bei Frequenzen, die von Null und unendlich verschieden sind, durch einen Koeffizienten der Schwingkreisrealisierung und unabhängig von allen anderen Koeffizienten festgelegt. Die Auswirkung einer Koeffizientenänderung auf die Empfindlichkeit des gesamten Dämpfungsverlaufs bleibt jedoch relativ gering. Daraus folgt, daß durch diskrete Optimierung Lösungen mit extrem kurzer einfacher Wortlänge der Multipliziererkoeffizienten gefunden wurden [16-18].

Im Gegensatz zu den Abzweig-Wellendigitalfiltern werden die Dämpfungspole in Brücken-Wellendigitalfilterstrukturen durch Kompensationseffekte, d.h., durch eine gegenseitige Auslöschung von Signalen, hervorgerufen. Daher sind die Brückenschaltungen im Sperrbereich sehr empfindlich. Das bedeutet, geringe Koeffizientenänderungen können zu einem unerwünschten Dämpfungsverhalten führen. Aus diesem Grunde erfordern die Brückenstrukturen in der Regel eine längere (in Festkommazahlen-Darstellung) Koeffizientenwortlänge, um die vorgeschriebene Dämpfungsanforderung erfüllen zu können.

Um die Abweichung der Sperrdämpfung abzuschätzen, wenn ein Multipliziererkoeffizient von seinem Nominalwert γ_0 abweicht, betrachten wir das in Bild 2.8 dargestellte symmetrisch, resistiv abgeschlossene Zweitor mit den kanonischen Impedanzen Z_1 und Z_2 und mit dem Last- und Quellwiderstand R . Die zu den kanonischen Impedanzen Z_1 bzw. Z_2 zugehörigen Reflektanzen S_1 und S_2 sind durch die folgenden Beziehungen gegeben:

$$S_1 = \frac{Z_1 - R}{Z_1 + R} \quad , \quad S_2 = \frac{Z_2 - R}{Z_2 + R} \quad . \quad (2.41)$$

Für die Reflektanz S_{11} und die Transmittanz S_{21} des Zweitores gilt der Zusammenhang:

$$S_{11} = \frac{B_1}{A_1} = \frac{S_2 + S_1}{2} \quad , \quad S_{21} = \frac{B_2}{A_1} = \frac{S_2 - S_1}{2} \quad . \quad (2.42)$$

Die Betriebsdämpfung α hängt mit der charakteristischen Funktion Ψ für reelle Frequenzen, d.h. $\psi = j\phi$, zusammen durch

$$\alpha(\phi) = 10 \log [1 + |\Psi(j\phi)|^2] \quad , \quad (2.43)$$

wobei $\Psi = \frac{S_{11}}{S_{21}} = \frac{Z_1 Z_2 - R^2}{(Z_2 - Z_1)R}$ ist.

In unserem Fall, d.h. für symmetrisches Reaktanzzweitor ($Z_1 = jx_1$ und $Z_2 = jx_2$), läßt sich $\Psi(j\phi)$ ausdrücken als

$$\Psi(j\phi) = j \frac{x_2(\phi)x_1(\phi) + R^2}{[x_2(\phi) - x_1(\phi)]R} \quad . \quad (2.44)$$

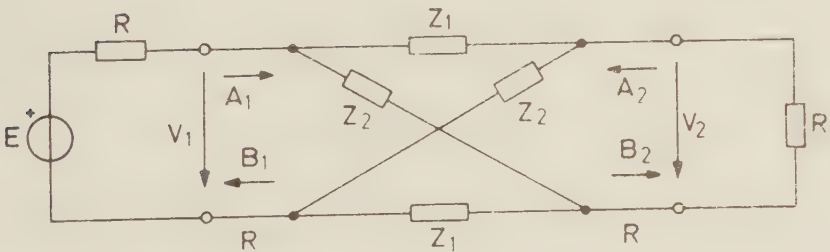


Bild 2.8: Symmetrisches Zweitor.

Wie man aus (2.42) ersehen kann, erhalten wir das in Bild 2.9 dargestellte Wellenflußdiagramm für ein Brücken-Wellendigitalfilter. Diese Realisierung der Reflektanz S_{11} und/oder der Transmittanz S_{21} ist einfach die Realisierung von den Reflektanzen S_1 und S_2 bzw. Z_1 und Z_2 , und dazu gibt es zahlreiche Möglichkeiten [20].

Weicht ein Multipliziererkoeffizient in einer der Brückenwei

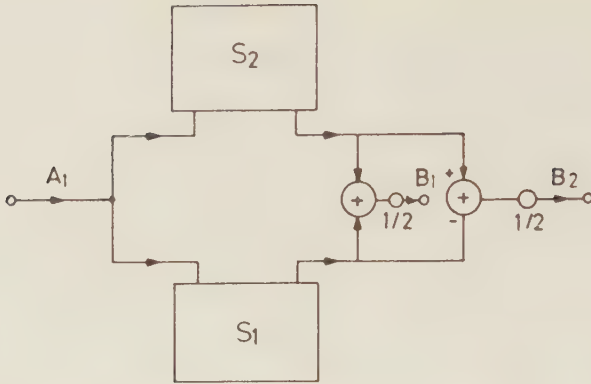


Bild 2.9: Das zur in Bild 2.8 dargestellten Schaltung zugehörige Wellenflußdiagramm.

ge von seinem Nominalwert ab, so ändert sich demzufolge der Wert der zugehörigen kanonischen Impedanz im Referenzfilter. ϕ_0 sei eine Frequenzstelle im Sperrbereich. Dann haben wir

$$|\Psi(j\phi_0)| \gg 1,$$

und die Betriebsdämpfung $\alpha(\phi_0)$ läßt sich ausdrücken als

$$\alpha(\phi_0) = 20 \log \left| \frac{x'_2(\phi_0)x'_1(\phi_0) + 1}{x'_2(\phi_0) - x'_1(\phi_0)} \right|, \quad (2.45)$$

wobei $x'_1(\phi_0) = \frac{x_2(\phi_0)}{R}$ und $x'_2(\phi_0) = \frac{x_1(\phi_0)}{R}$ ist.

Wenn $x'_1(\phi_0)$ und $x'_2(\phi_0)$ von ihrem Nominalwert um $\Delta x'_1(\phi_0)$ und $\Delta x'_2(\phi_0)$ abweichen, dann nimmt die Betriebsdämpfung den neuen Wert

$$\alpha(\phi_0) + \Delta\alpha(\phi_0) = 20 \log \left| \frac{[x'_2(\phi_0) + \Delta x'_2(\phi_0)][x'_1(\phi_0) + \Delta x'_1(\phi_0)] + 1}{x'_2(\phi_0) - x'_1(\phi_0) + \Delta x'_2(\phi_0) - \Delta x'_1(\phi_0)} \right| \quad (2.46)$$

Unter der Voraussetzung, daß $\Delta x'_1(\phi_0) \ll x'_1(\phi_0)$ und $\Delta x'_2(\phi_0) \ll x'_2(\phi_0)$ ist, gilt die folgende Beziehung

$$\Delta\alpha(\phi_0) = 20 \log \left| \frac{1}{1 + \frac{\Delta x'_1(\phi_0) - \Delta x'_2(\phi_0)}{x'_1(\phi_0) - x'_2(\phi_0)}} \right|. \quad (2.47)$$

Will man im Sperrbereich eine größere Dämpfung, so muß die Diffe-

renz $x'_2(\phi_0) - x'_1(\phi_0)$ möglichst klein sein. Als Beispiel nehmen wir an

$$x'_1(\phi_0) = 1 \quad , \quad x'_2(\phi_0) = 0,999 \quad .$$

Dann läßt sich die Betriebsdämpfung $\alpha(\phi_0)$ durch (2.45) berechnen als

$$\alpha(\phi_0) \cong 66 \text{ dB} \quad .$$

Weicht nur $x'_1(\phi_0)$ von seinem nominellen Wert um 1 % ab, so ergibt die Abweichung der Dämpfung $\Delta\alpha(\phi_0)$ durch (2.47) unter denselben vorausgesetzten Werten von $x'_1(\phi_0)$ und $x'_2(\phi_0)$ folgendes

$$\Delta\alpha(\phi_0) \cong -20,8 \text{ dB} \quad .$$

Das bedeutet, die Dämpfung $\alpha(\phi_0)$ nimmt statt 66 dB, wenn alle Bauelemente im Filter ihren nominellen Wert annehmen, den neuen Wert 45,2 dB an, wenn die kanonische Impedanz $x_1(\phi_0)$ von ihrem Nominalwert um 1 % abweicht. Aus Gleichung (2.47) ist zu ersehen, daß die Abweichung der kanonischen Impedanz $x_1(\phi_0)$ und/oder $x_2(\phi_0)$ zu nicht tolerierbarem Dämpfungsverlauf im Sperrbereich führen kann. Dieselbe Aussage gilt auch bei den Polstellen. Um dies zu erklären, nehmen wir an, daß ϕ_∞ die Frequenz einer Polstelle sei. Dann gilt also $\alpha(\phi_\infty) = \infty$, d.h., $x'_2(\phi_\infty) = x'_1(\phi_\infty)$, wenn alle Bauelemente ihren Nominalwert annehmen. Weichen $x'_1(\phi_\infty)$ und $x'_2(\phi_\infty)$ ab, so nimmt die Dämpfung den neuen Wert $\alpha'(\phi_\infty)$ an:

$$\alpha'(\phi_\infty) = 20 \log \left| \frac{x_1'^2(\phi_\infty) + 1}{\Delta x_2'(\phi_\infty) - \Delta x_1'(\phi_\infty)} \right| \quad , \quad (2.48)$$

unter der Voraussetzung, daß $|\Delta x_1| \ll x_1$ und $|\Delta x_2| \ll x_2$ ist.

Aus diesem Grunde erfordern Brücken-Wellendigitalfilter im allgemeinen eine längere Koeffizientenwortlänge in Festkommazahlen-Darstellung, um vorgeschriebene Dämpfungsanforderungen erfüllen zu können. Trotz dieser Nachteile für Brückenfilter lassen sich Lösungen mit sehr einfachen im CSD-Code dargestellten Koeffizienten angeben, die nur mit geringem Aufwand zu realisieren sind [20,61].

2.4 Abhängigkeit zwischen der Empfindlichkeit und dem Quantisierungsrauschen

Die Empfindlichkeit $\frac{\partial \alpha'}{\partial \gamma_i}$ hat offensichtlich wesentlichen

Einfluß auf das Signal-Rausch-Verhältnis am Ausgang eines Filters [12,21-23,72]. Wir wollen diese Aussage quantitativ machen.

Zunächst betrachten wir ein Wellendigitalfilter (Bild 2.10) mit einem Multiplizierer zwischen den Knoten 3 und 4.

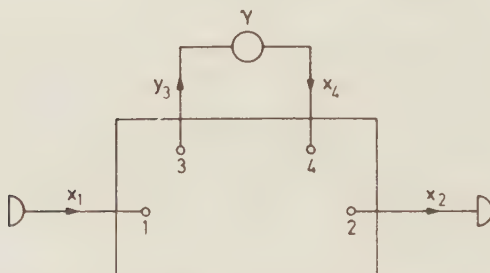


Bild 2.10: Ein Digitalfilter mit einem Multiplizierer .

Die Transmittanz vom Knoten 1 zum Knoten 2 wird gegeben als

$$S'_{21}(j\omega) = y_2/x_1 = e^{-\Gamma'(j\omega)} \quad , \quad (2.49a)$$

$$\Gamma'(j\omega) = \alpha'(\omega) + j\beta'(\omega) \quad . \quad (2.49b)$$

Wir nehmen an, daß der Multiplizierer mit Gleitkommaarithmetik realisiert ist. In der Arbeit [13] ist bewiesen worden, daß das Rausch-Signal-Verhältnis (NSR) oder der Dynamikbereich [24] von der Empfindlichkeit abhängig ist. Diese Abhängigkeit lautet:

$$NSR(\omega_o) = K(\omega_o) \left| \frac{\partial \alpha'(\omega_o)}{\partial \ln \gamma} + j \frac{2\phi_o}{\pi} \int_0^\infty \frac{\partial \alpha'(\phi)}{\partial \ln \gamma} \frac{d\phi}{\phi^2 - \phi_o^2} \right| \quad , \quad (2.50)$$

wobei ω_o die Frequenz des Eingangssignals,

$$\psi = \tan(\omega T/2) \quad \text{und} \quad \phi_o = \tan(\omega_o T/2) \quad \text{ist.}$$

$K(\omega_o)$ ist eine Konstante im Fall der Gleitkommaarithmetik.

Der Gleichung (2.50) entnehme man, daß die geringe Empfindlichkeit $\frac{\partial \alpha}{\partial \gamma}$ zu geringem Quantisierungsrauschen führt. Im Digitalfilterbereich machen sich die Wellendigitalfilter diese Eigenschaften zunutze.

2.5 Stabilitätsverhalten der Wellendigitalfilter

Das Stabilitätsverhalten der Wellendigitalfilter ist der Hauptgrund ihrer Bedeutung. Ein direkter Beweis der linearen Stabilität eines Wellendigitalfilters wurde möglich, nachdem der klassische Leistungse- und Passivitätsbegriff auf den Wellendigitalfilterbereich übertragen wurde [25]. Da bei einem Digitalfilter die Leistung keine physikalische Bedeutung mehr hat, sondern eine mathematische Formulierung darstellt, wurde der Begriff der Pseudoleistung eingeführt. Mit Hilfe der Pseudoleistung wurden die Ausdrücke Pseudopassivität und Pseudoverlustfreiheit definiert, die in einem Wellendigitalfilter eine ähnliche Rolle spielen wie die Passivität und die Verlustfreiheit in einem Referenzfilter.

Bei rekursiven Digitalfiltern, zu denen Wellendigitalfilter gehören, laufen im Filter berechnete Größen beim nächsten Iterationsschritt wieder in das Rechenwerk des Filters zurück. Will man die Rechenoperationen innerhalb des Digitalfilters ideal linear durchführen, so muß mit jedem Rechenzyklus die Signalwortlänge erhöht werden. Diese Erhöhung ergibt sich hauptsächlich aus den Multiplikationen im Fall der Festkommaarithmetik oder aus den Multiplikationen und den Additionen bei der Gleitkommaarithmetik. Aus Aufwandsgründen müssen Quantisierungsoperationen (Schneide- oder Rundungsoperationen) durchgeführt werden, die die Signalwortlänge wieder auf die ursprüngliche Wortlänge reduzieren, damit die Anzahl der notwendigen Binärstellen nicht ins Unendliche wächst.

Infolge der Quantisierungsoperationen, die nicht linear sind, treten in einem rekursiven Digitalfilter nichtlineare Stabilitätsprobleme auf, die sich aber auch wegen der möglichen Überläufe, d.h. Überschreitungen des zur Darstellung der Signale verfügbaren Zahlenbereiches, ergeben.

Bei Wellendigitalfiltern trägt das Pseudopassivitätskonzept aber auch zur Lösung der nichtlinearen Stabilitätsprobleme bei. Zur Erhaltung der Pseudopassivität eines Wellendigitalfilters, welches

unter Berücksichtigung der endlichen Signalwortlänge arbeitet, wurde bei abgeschaltetem Eingangssignal eine Bedingung gegeben [26]. Diese Pseudopassivitätsbedingung gilt ja für alle möglichen Zahlendarstellungen (z.B., Festkomma-, Gleitkommadarstellung etc.). Es sei an dieser Stelle ausdrücklich darauf hingewiesen, daß die in [26] gegebene Pseudopassivitätsbedingung hinreichend und nicht notwendig ist. Neben der Stabilität bei abgeschaltetem Eingangssignal ist insbesondere auch die sogenannte Antwortstabilität eines Digitalfilters von Bedeutung. Unter Antwortstabilität versteht man, daß die Störungen des Ausgangssignals, die durch einen einmal aufgetretenen Überlauf verursacht werden, auch bei anliegendem Eingangssignal wieder abklingen [28].

Im Fall der Festkommarealisierung eines Wellendigitalfilters konnten technisch einfach zu realisierende Maßnahmen angegeben werden, um nach dem Abschalten des Eingangssignals sämtliche parasitäre Schwingungen zu unterdrücken [26]. Diese Unterdrückung von parasitären Schwingungen kann sogar erreicht werden, wenn das Wellendigitalfilter in einem größeren System innerhalb einer Schleife arbeitet, die durch die Zweidraht/Vierdraht-Nachrichtenverbindungen entsteht [50]. Die Verwirklichung einer Schleifenstabilität wurde nur für Wellendigitalfilter sowie für nichtrekursive Digitalfilter nachgewiesen. Darüber hinaus konnte mit dem Konzept der inkrementalen Pseudopassivität gewährleistet werden, daß bei Verwendung geeigneter Überlaufcharakteristiken an den Adaptorausgängen die Wellendigitalfilter nämlich inkremental pseudopassiv sind [29,30].

Es sei an dieser Stelle darauf hingewiesen, daß die inkrementalen Pseudopassivitätseigenschaften eines Wellendigitalfilters nicht nur die Antwortstabilität beinhalten, sondern auch die Stetigkeit des Übertragungsverhaltens eines Filters. Das bedeutet, daß kleine Änderungen des Eingangssignals nur kleine Änderungen des Ausgangssignals verursachen können, und allgemeiner, daß geringe Übersteuerungen am Eingang nur geringfügige Verzerrungen am Ausgang des Filters zur Folge haben können. Es hat sich also gezeigt, daß sich beim Wellendigitalfilter, in dem die Signalparameter als Festkommagrößen dargestellt sind, die nichtlinearen Stabilitätsprobleme vollständig lösen lassen.

Bei Verwendung der Zweierkomplement-Gleitkomma-Arithmetik kann auch die Pseudopassivitätsbedingung eingehalten werden. Weiterhin konnte mit dem Konzept der inkrementalen Pseudopassivität nachgewie-

sen werden, daß Gleitkomma-Wellendigitalfilter antwortstabil sind. Die Maßnahmen werden im Kapitel 4 ausführlich diskutiert. Durch diese interessanten Stabilitätseigenschaften sind die Wellendigitalfilter den anderen rekursiven Digitalfiltern überlegen. Daher können beim Einsatz von sowohl Festkomma- als auch Gleitkomma-Wellendigitalfiltern keine Stabilitätsprobleme verursacht werden, da sich diese Filter bei geeigneten Maßnahmen voll wie passive Schaltungen verhalten.

3. Gleitkommazahlendarstellung

3.1 Vorbemerkungen

Aus Aufwandsgründen können in einem Digitalfilter nur endlich viele unterschiedliche rationale Zahlen verarbeitet werden. Für die Darstellung dieser Zahlen kommen verschiedene Zahlensysteme und Kommastellungen in Betracht. Die Wahl der Darstellung hängt von der Realisierungsart des Filters ab. In dieser Arbeit werden wir uns auf den Fall der Darstellung einer Zahl als Gleitkommagröße beschränken.

Eine endliche reelle Zahl x_0 wird in einem Potenzzahlensystem der Basis B durch eine allgemeine unendliche Reihe der Form

$$x_0 = \pm \sum_{i=n}^{-\infty} \alpha_i B^i \quad (3.1)$$

dargestellt, wobei B eine ganze Zahl ist und α_i die Bedingung

$$0 \leq \alpha_i \leq B-1 \quad (3.2)$$

erfüllt. Das Komma trennt die Koeffizienten der Reihenglieder, die zu B^0 und B^{-1} proportional sind.

Wie man aus (3.1) ersehen kann, läßt sich das Komma an eine andere beliebige Stelle verschieben, falls man nachträglich mit einer geeigneten Potenz der Basis B multipliziert. Schiebt man das Komma links vor die erste von Null verschiedene Ziffer, so spricht man von normalisierter Gleitkommadarstellung der Zahl x_0 . In diesem Fall läßt sich die reelle Zahl x_0 in der Form

$$x_0 = m_0 \cdot B^e \quad (3.3)$$

darstellen. Dabei bezeichnet man die Bruchzahl

$$m_0 = \pm \sum_{i=-1}^{-\infty} \alpha_i B^i \quad (3.4)$$

als die Mantisse und die ganze Zahl e als den Exponenten der Darstellung (3.3). Die Mantisse genügt der Bedingung

$$1/B \leq |m_0| < 1 \quad (3.5)$$

Die sogenannte normalisierte Gleitkommazahlendarstellung mit endlicher Wortlänge für die Mantisse und für den Exponenten liegt der Darstellung einer reellen Zahl sowohl in Rechenanlagen als auch in

einem Digitalfilter zugrunde. Daher wird die unendliche Reihe (3.4) an einer geeigneten Stelle $i=t_{\text{man}}$ abgebrochen, so daß x_0 nur näherungsweise dargestellt werden kann, was durch x angedeutet wird. Für die Zahl x benutzt man die abgekürzte Schreibweise

$$x = [x_0]_q = m \cdot B^e \quad (3.6)$$

Damit läßt sich m durch

$$m = \pm \sum_{i=-1}^{-t_{\text{man}}} \alpha_i B^i \quad \text{mit} \quad \alpha_i \in \{0, 1, \dots, B-1\} \quad (3.7)$$

ausdrücken, wobei t_{man} die Mantissenwortlänge bezeichnet.

Aus der vorstehenden Definition (3.6) geht unmittelbar hervor, daß für ein gegebenes Gleitkommasystem GS vier Konstanten bestehen, nämlich B , t_{man} , e_{min} und e_{max} . Dabei bezeichnet e_{max} den Maximal-exponenten und e_{min} den Minimal-exponenten. Ferner entnehme man dieser Definition, daß es endlich viele Gleitkommazahlen der in (3.6) gegebenen Form gibt. Diese Zahlen sind zwischen der größten und kleinsten Gleitkommazahl ungleichmäßig verteilt. Die Verteilung ist natürlich symmetrisch zum Nullpunkt [31,32].

Das Zahlensystem mit $B=2$, also das Binär- oder Dualzahlensystem, hat bisher große Bedeutung erlangt. Der Grund besteht darin, daß die bisher vorwiegend verfügbaren Digitalschaltungen Binärschaltkreisfamilien sind. Daher wird weiterhin in dieser Arbeit das Binärzahlensystem behandelt.

In diesem Fall läßt sich die Zahl x in der Form

$$x = m \cdot 2^e \quad (3.8)$$

schreiben, wobei m ein echter Dualbruch ist und der Bedingung

$$1/2 \leq |m| < 1 \quad (3.9)$$

genügt.

Der Exponent e ist eine ganze Zahl der Form

$$e = \sum_{i=t_{\text{exp}}-1}^0 \alpha_i 2^i \quad (3.10)$$

Hier bezeichnet t_{exp} die Exponentenwortlänge.

An dieser Stelle ist es sinnvoll, die Wahrscheinlichkeitsdich-

tefunktion einer in GS beliebig liegenden Gleitkommazahl zu berechnen.

3.2 Wahrscheinlichkeitsdichtefunktion einer Gleitkommazahl

Wir nehmen an, daß die Mantisse m in dem gegebenen Bereich (3.9) und auch der Exponent e in dem Bereich

$$-e_{\max} \leq e \leq e_{\max} \quad (3.11)$$

gleichverteilt sind. Ferner wird angenommen, daß keine statistische Abhängigkeit zwischen der Mantisse und dem Exponenten besteht. Bild 3.1 zeigt die Verteilung der Wahrscheinlichkeitsdichtefunktion der Mantisse m . $P(m)$ läßt sich als

$$P(m) = 1 \quad \text{für} \quad \frac{1}{2} \leq |m| < 1 \quad (3.12)$$

bestimmen. Die Varianz σ_m^2 ist gegeben durch

$$\begin{aligned} \sigma_m^2 &= E[m^2] - E^2[m] = E[m^2] \\ &= 7/12 \end{aligned} \quad (3.13)$$

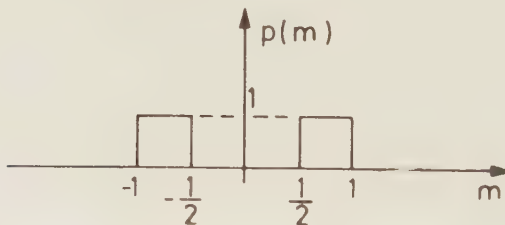


Bild 3.1: Verteilung der Wahrscheinlichkeitsdichtefunktion $P(m)$.

Bild 3.2 zeigt die Verteilung der Wahrscheinlichkeitsdichtefunktion des Exponenten.

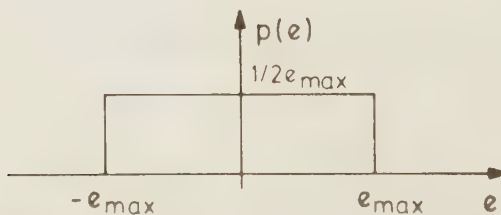


Bild 3.2: Die Verteilung der Wahrscheinlichkeitsdichtefunktion $P(e)$.

Auf ähnliche Weise lassen sich $P(e)$ und die Varianz σ_e^2 schreiben

$$P(e) = \frac{1}{2 e_{\max}} \quad \text{für } |e| \leq e_{\max} \quad , \quad (3.14a)$$

$$\sigma_e^2 = \frac{e_{\max}^2}{3} \quad . \quad (3.14b)$$

Man kann davon ausgehen, daß eine Gleitkommazahl x gemäß (3.8) in folgender Form geschrieben werden kann

$$x = m \cdot y \quad , \quad (3.15)$$

wobei $y=2^e$ ist. Unter der oben angeführten Annahme, daß keine statistische Abhängigkeit zwischen der Mantisse und dem Exponenten besteht, läßt sich die Wahrscheinlichkeitsdichtefunktion $P(x)$ schreiben als

$$P(x) = P(m) \cdot P(y) \quad . \quad (3.16)$$

Nun bestimmen wir $P(y)$. Bild 3.3 zeigt die Relation $y=2^e$.

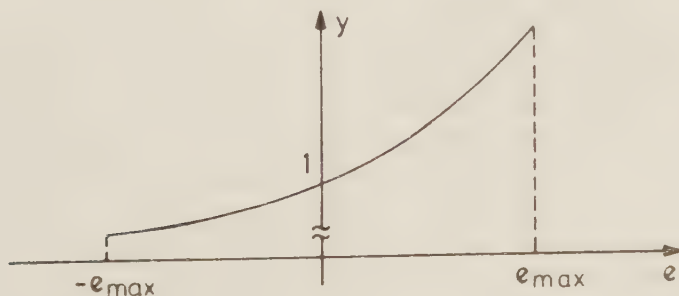


Bild 3.3: Die Relation $y=2^e$.

Dem Bild 3.3 entnimmt man die folgende Beziehung

$$P(y) |dy| = P(e) |de| \quad . \quad (3.17)$$

Daraus folgt, daß sich die Wahrscheinlichkeitsdichtefunktion $P(y)$ wie folgt schreiben läßt:

$$P(y) = \frac{P(e)}{2^e \cdot \ln 2} \quad (3.18)$$

Setzen wir (3.12) und (3.18) in (3.16) mit der Verwendung von (3.14a) und (3.8), so erhalten wir

$$P(x) = \frac{1}{[2 e_{\max} \cdot \ln 2] 2^e} = \frac{m}{[2 e_{\max} \cdot \ln 2] x} \quad (3.19)$$

Wie man aus der Gleichung (3.19) ersehen kann, nimmt $P(x)$ mit wachsendem Exponenten ab.

3.3 Quantisierungsfehler bei Gleitkommazahlen

Die Begrenzung der Mantissenwortlänge t_{man} bei Gleitkommazahendarstellung wird Quantisierung genannt. Es gilt für die Gleitkommaquantisierungsstufe Q_g

$$Q_g = 2^{-t_{\text{man}}} 2^e = 2^{e-t_{\text{man}}} \quad (3.20)$$

Im Fall der Festkommadarstellung ist die Quantisierungsstufe Q_f gleichmäßig und hat den Wert

$$Q_f = 2^{-t} \quad , \quad (3.21)$$

wobei t die Festkommawortlänge angibt. Bild 3.4 zeigt die Gleit- und die Festkommaquantisierungsstufe, wenn $t=t_{\text{man}}$ ist.

Dem Bild 3.4 entnimmt man, daß die Quantisierungsstufe von der Gleitkommazahl abhängig ist, wohingegen sie bei Festkommaquantisierung eine Konstante bleibt. Dies bedeutet, daß die durch Gleitkommaquantisierung entstehenden Fehler für kleinere Größen klein sind.

Um diese Aussage zu erläutern nehmen wir an, daß der Exponent einer Gleitkommazahl ohne Quantisierungsfehler dargestellt wird, während die Mantisse eine endliche Stellenzahl t_{man} besitzt. Der durch die Quantisierung entstehende absolute Fehler wird definiert durch

$$\epsilon_{\text{abs}} = x_o - [x_o]_q \quad (3.22)$$

Es ist plausibel, daß dieser absolute Fehler in dem Bereich, wo sich der Exponent nicht ändert, gleich verteilt ist. Tabelle 3.1

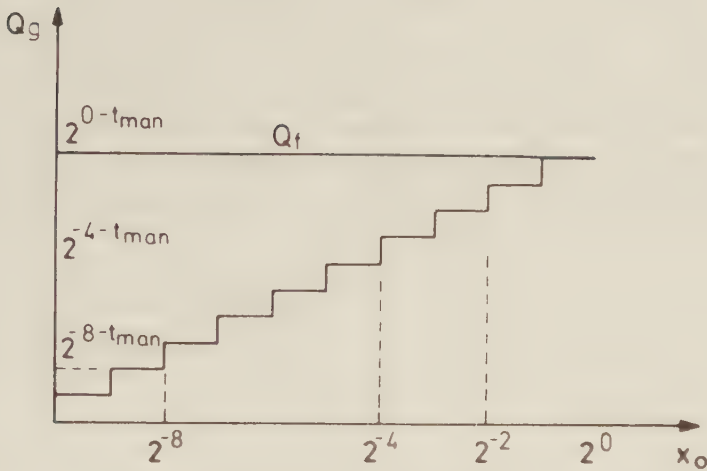


Bild 3.4 Quantisierungsstufen Q_g und Q_f .

enthält die Zusammenstellung der Schranken für die absoluten Quantisierungsfehler bei Gleitkommazahlendarstellung.

	Absoluter Fehler ϵ_{abs} bei der Gleitkommazahlendarstellung
Rundung	$-\frac{Q_g}{2} \leq \epsilon_{\text{abs}} \leq \frac{Q_g}{2}$
Betragsschneiden	$0 \leq \epsilon_{\text{abs}} < Q_g$ für $x_0 > 0$ $-Q_g < \epsilon_{\text{abs}} \leq 0$ für $x_0 < 0$
Zweierkomplement-Schneiden	$0 \leq \epsilon_{\text{abs}} < Q_g$

Tabelle 3.1: Schranken der absoluten Quantisierungsfehler bei Gleitkommazahlendarstellung.

Der größte Vorteil einer Gleitkommaarithmetik besteht darin, daß der relative Quantisierungsfehler im Gegensatz zur Festkommaarithmetik immer auch für kleinere Zahlen klein bleibt. Der relative Fehler ist definiert durch

$$\epsilon_r = \frac{\epsilon_{abs}}{x_0} = \frac{x_0 - [x_0]_q}{x_0} = \frac{m_0 - m}{m_0} \quad (3.23)$$

Tabelle 3.2 enthält die Schranken für die relativen Quantisierungsfehler bei Gleitkommadarstellung. Die Bilder 3.5 bis 3.7 zeigen die Quantisierungskennlinien für Rundung, Betragsschneiden und Zweierkomplementschneiden bei Gleitkommazahlendarstellung sowie die Kurven des absoluten und relativen Fehlers für $t_{man}=3$ Bits und $t_{exp}=1$ bit jeweils ohne Vorzeichenbit.

	Relativer Fehler ϵ_r bei der Gleitkommazahlendarstellung
Rundung	$-2^{-t_{man}} < \epsilon_r < 2^{-t_{man}}$
Betragsschneiden	$0 < \epsilon_r < 2^{-t_{man}+1}$
Zweierkomplement-Schneiden	$0 < \epsilon_r < 2^{-t_{man}+1} \quad \text{für } x_0 > 0$ $-2^{-t_{man}+1} < \epsilon_r \leq 0 \quad \text{für } x_0 < 0$

Tabelle 3.2: Schranken des relativen Quantisierungsfehlers bei Gleitkommazahlendarstellung.

3.4 Verteilungsdichtefunktion des relativen Quantisierungsfehlers

Es ist sinnvoll, den relativen Quantisierungsfehler durch seine Verteilungsdichtefunktion zu beschreiben, da er für die Berechnung des Signalrauschverhältnisses in einem Gleitkomma-Digitalfilter Verwendung findet.

Es wurde gezeigt [33-35], daß unterschiedliche Verteilungen der Mantisse oder des relativen Fehlers bei der Berechnung der Varianz $\sigma_{\epsilon_r}^2$ auf unterschiedliche Werte führen. In [34] ist die Wahrscheinlichkeitsdichtefunktion $P(\epsilon_r)$ des relativen Fehlers im Fall der Rundung und des Schneidens für unterschiedliche Verteilungen der Mantisse m gegeben. Im weiteren wird eine in ihrem Bereich nach (3.9) gleichverteilte Mantisse zugrunde gelegt. Bild 3.8 zeigt die Wahrscheinlichkeitsdichtefunktion des relativen Fehlers ϵ_r bei Gleitkommaquantisierung durch Runden, Betragsschneiden und Zweier-

komplementschneiden. Tabelle 3.3 enthält den Erwartungswert und die Varianz der jeweiligen Wahrscheinlichkeitsdichtefunktion[40].

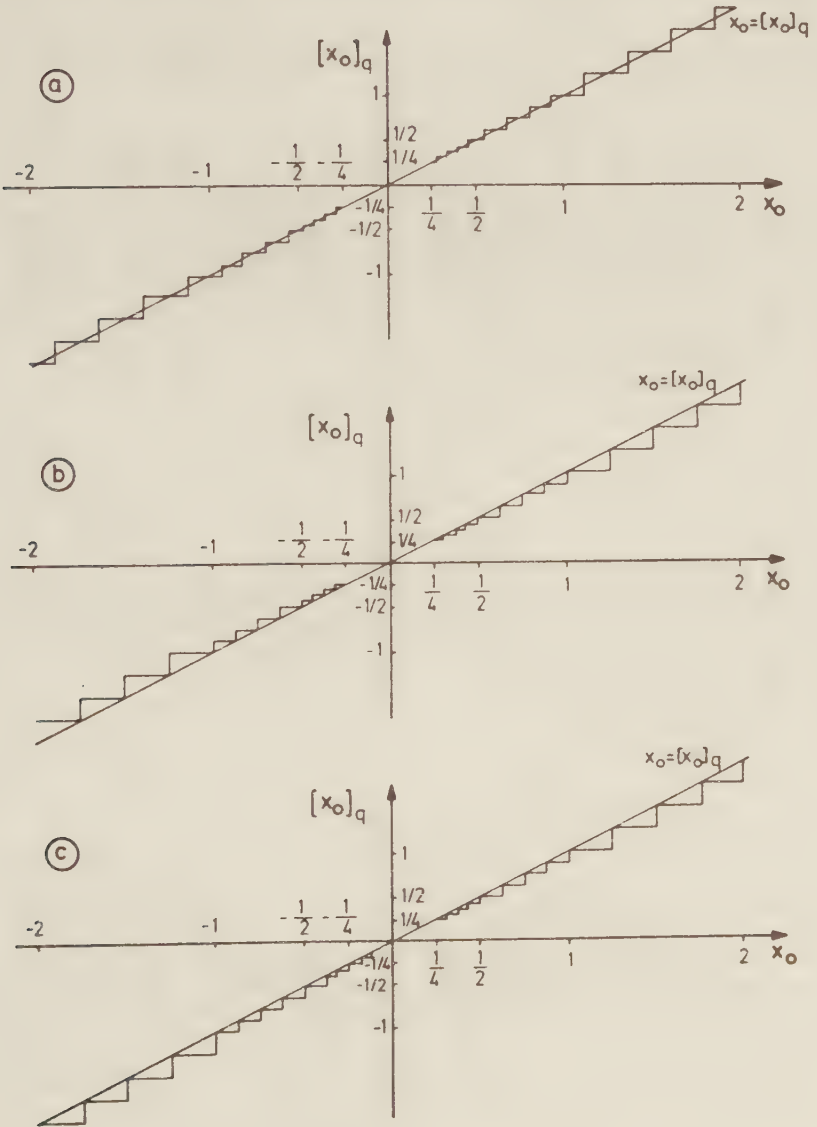


Bild 3.5: Quantisierungskennlinien für die Quantisierung von Gleitkommazahlen durch a) Runden, b) Betragsschneiden und c) Zweierkomplement-Schneiden ($t_{\text{man}} = 3$ Bits).

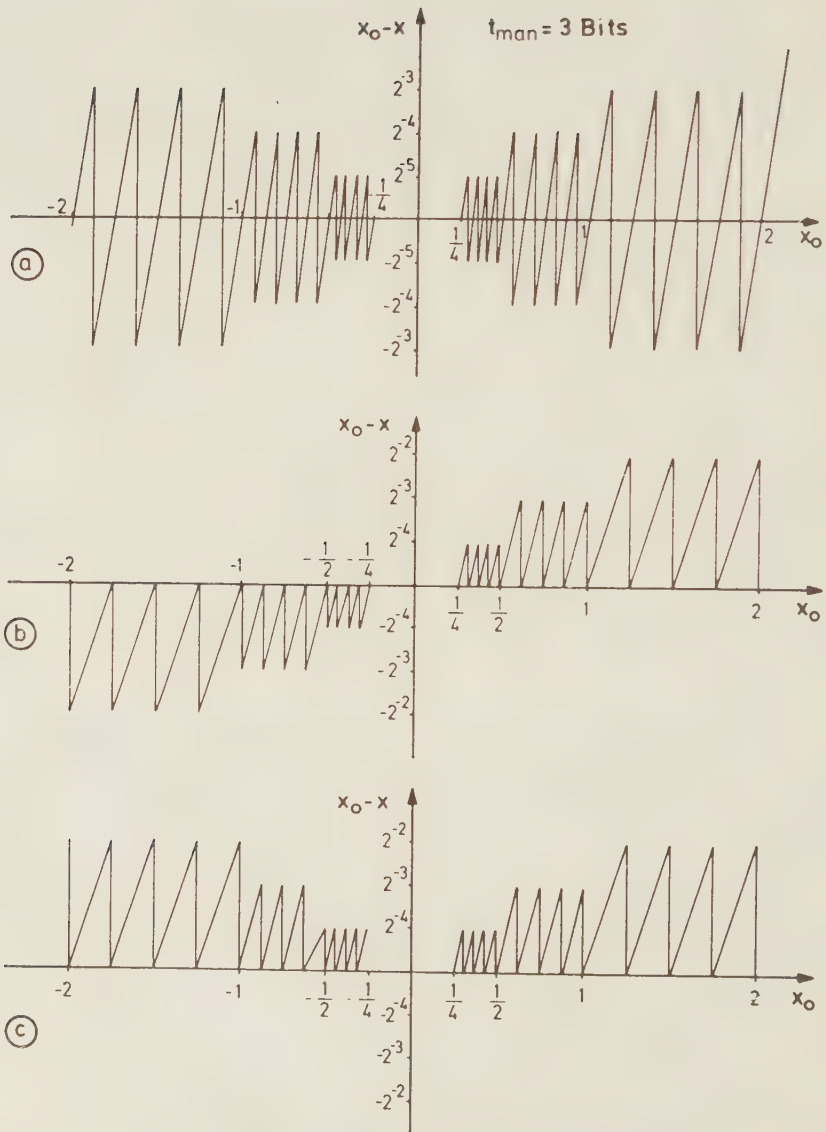


Bild 3.6: Kurven des absoluten Fehlers für a) Runden, b) Betragschneiden und c) Zweierkomplementschneiden.

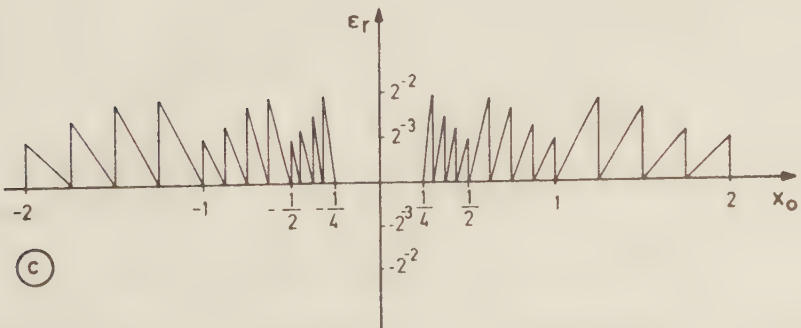
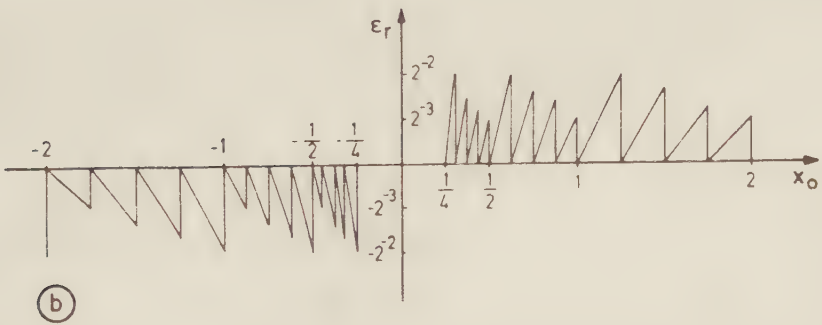
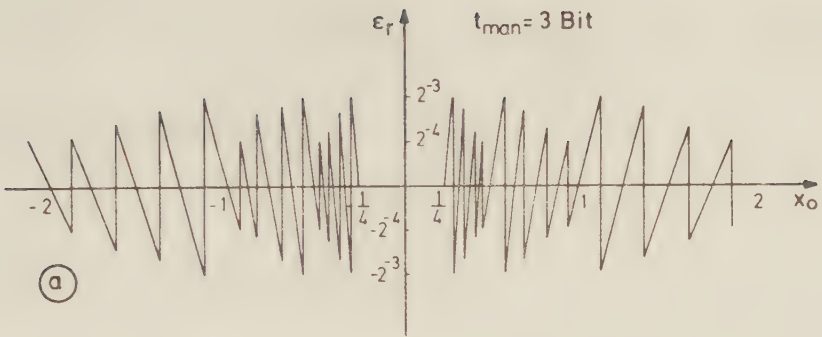


Bild 3.7: Kurven des relativen Fehlers für a) Runden, b) Zweierkomplementerschneiden und c) Betragsschneiden.

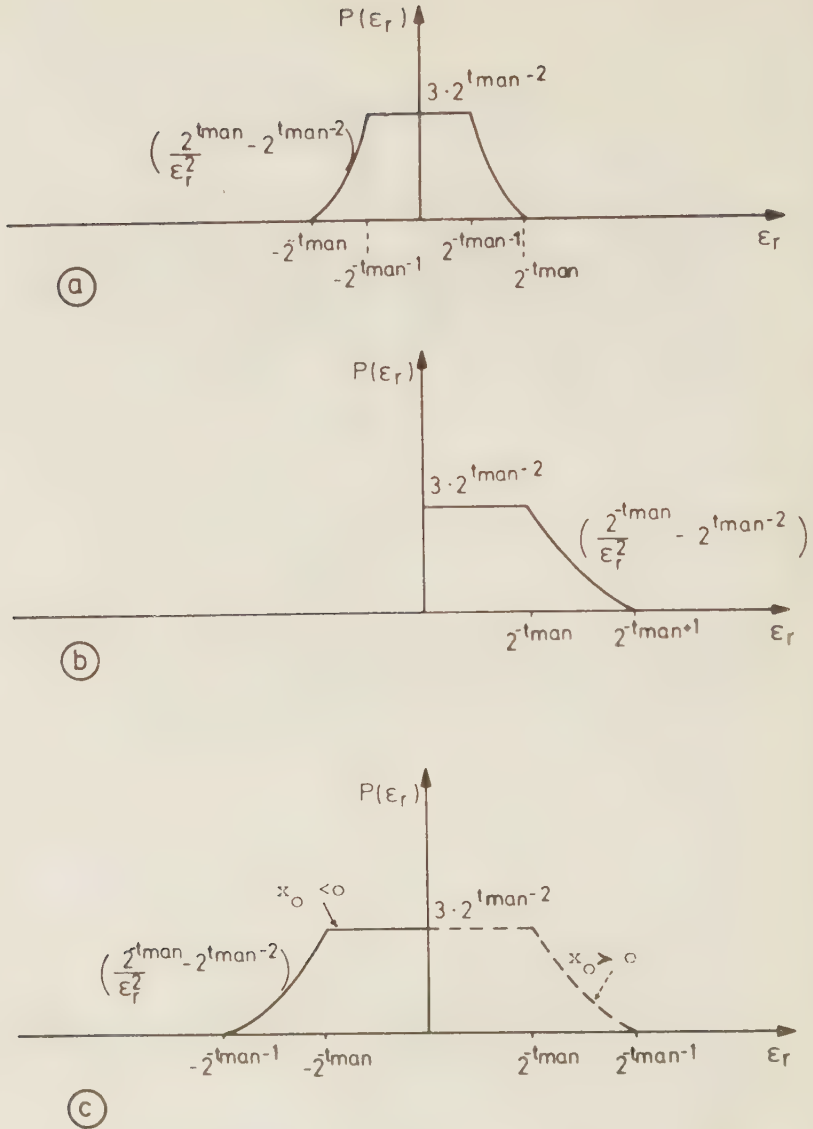


Bild 3.8: Wahrscheinlichkeitsdichtefunktion des relativen Fehlers bei Gleitkommaquantisierung durch a) Runden, b) Betragschneiden, c) Zweierkomplementschneiden.

	Erwartungswert	Varianz $\sigma_{\epsilon_r}^2$
Rundung	0	$(1/6) 2^{-2t_{\text{man}}}$
Betragserschneiden	$2^{-t_{\text{man}}} \ln 2$	$(2/3 - \ln^2 2) 2^{-2t_{\text{man}}}$
Zweierkomplement- Schneiden	$2^{-t_{\text{man}}} \ln 2$ für $x_0 > 0$ $-2^{-t_{\text{man}}} \ln 2$ für $x_0 < 0$	$(2/3 - \ln^2 2) 2^{-2t_{\text{man}}}$

Tabelle 3.3: Erwartungswert und Varianz des relativen Fehlers bei Gleitkommaquantisierung.

3.5 Wahrscheinlichkeitsdichtefunktion des absoluten Fehlers ϵ_{abs}

Es ist zweckmäßig, den absoluten Quantisierungsfehler durch seine Wahrscheinlichkeitsdichtefunktion zu beschreiben. Im allgemeinen kann der absolute Fehler ϵ_{abs} auf folgende Weise geschrieben werden:

$$\epsilon_{\text{abs}} = \epsilon \cdot 2^e = \epsilon \cdot y, \quad (3.24)$$

wobei ϵ im Fall der Rundung durch

$$-2^{-t_{\text{man}}-1} \leq \epsilon \leq 2^{-t_{\text{man}}-1} \quad (3.25)$$

gegeben ist. Unter der Annahme, daß keine statistische Abhängigkeit zwischen ϵ und e besteht, und ϵ in dem gegebenen Bereich (3.25) gleich verteilt ist, läßt sich die Wahrscheinlichkeitsdichte $P(\epsilon_{\text{abs}})$ ausdrücken als

$$P(\epsilon_{\text{abs}}) = P(\epsilon) P(y) \quad \text{und} \quad P(\epsilon) = 2^{-t_{\text{man}}}. \quad (3.26a)$$

Verwendet man (3.18) und (3.14a), so erhält man

$$P(\epsilon_{\text{abs}}) = 2^{-t_{\text{man}}} \cdot \epsilon / (2e_{\text{max}} (\ln 2) \epsilon_{\text{abs}}). \quad (3.26b)$$

Auf ähnliche Weise ist $P(\epsilon_{\text{abs}})$ für das Betrags- und Zweierkomplementschneiden durch (3.26b) in dem nach Tabelle 3.1 gegebenen Bereich für ϵ_{abs} gegeben. Wie man aus der Gleichung (3.26b) ersehen

kann, nimmt die Wahrscheinlichkeitsdichtefunktion $P(\epsilon_{\text{abs}})$ mit wachsendem Exponenten ab.

3.6 Fehler durch Begrenzung des Zahlenbereichs

Wie schon im vorigen Unterabschnitt 3.1 erwähnt wurde, besteht der Exponent einer Gleitkommazahlendarstellung aus endlicher Wortlänge, t_{exp} . Durch diese Begrenzung des Exponentenbereichs ergibt sich eine Begrenzung des Zahlenbereichs. Dies führt zu einer anderen Kategorie von Fehlern, je nachdem ob der Zahlenbereich überschritten oder unterschritten wird.

3.6.1 Exponentenüberlauf

Je nach der Darstellungsart der Mantisse läßt sich die betragsmäßig größte darstellbare Zahl im Fall der Vorzeichenbetragsdarstellung durch

$$|x_{\text{max}}| = (1 - 2^{-t_{\text{man}}}) 2^{e_{\text{max}}} \quad (3.27)$$

ausdrücken. Bei einer Zweierkomplementdarstellung erhalten wir

$$x_{\text{max}} = (1 - 2^{-t_{\text{man}}}) 2^{e_{\text{max}}}, \quad \text{für } x > 0 \quad (3.28a)$$

$$|x_{\text{max}}| = 2^{e_{\text{max}}} \quad \text{für } x < 0 \quad (3.28b)$$

Bei einer Überschreitung des Zahlenbereichs über die größte darstellbare Gleitkommazahl gemäß (3.27) oder (3.28) spricht man von Exponentenüberlauf (Überlauf). In diesem Fall wird üblicherweise die Sättigungscharakteristik verwendet [36,37], d.h., überlaufende Größen werden durch maximale (betragsmäßig) darstellbare Zahlen ersetzt.

3.6.2 Exponentenunterlauf

Die Unterschreitung des Zahlenbereichs unter die kleinste darstellbare Gleitkommazahl

$$|x_{\text{min}}| = 2^{e_{\text{min}} - 1} \quad (3.29)$$

wird Exponentenunterlauf oder einfach Unterlauf genannt. Treten während einer Rechnung Zahlen auf, welche betragsmäßig kleiner als die

Zahl (3.29) sind, so werden sie einfach durch Null ersetzt [32,36,37], wenn man den Verlust aller Ziffern in der Mantisse in Kauf nehmen will. Eine andere Methode zur Behandlung von Exponentenunterlauf ist gegeben [38]. Diese Behandlungsweise lautet: Wenn der Exponent e einer Gleitkommazahl gleich dem kleinsten darstellbaren Exponenten $e_{\min}$ ist, dann wird automatisch auf die Normierung verzichtet. In diesem Fall ist die kleinste darstellbare Zahl gegeben durch

$$|x_{\min}| = 2^{e_{\min} - t_{\max}} \quad (3.30)$$

3.7 Zahlenbereich und die Genauigkeit von Gleit- und Festkomma- darstellung

Der vorigen Diskussion kann man entnehmen, daß der Zahlenbereich einer Gleitkommazahlendarstellung mit der Wortlänge des Exponenten zunimmt. Um dies näher zu erläutern definiert man den Zahlenbereich einer Zahlendarstellung als den Abstand in Dezibel zwischen der größten und der kleinsten darstellbaren Zahl durch [40]

$$ZB = 20 \log \left| \frac{x_{\max}}{x_{\min}} \right| \quad (3.31)$$

Nun betrachten wir den Fall, wenn der Exponent in Vorzeichenbetrag dargestellt ist. Der Zahlenbereich für die Gleitkommadarstellung ergibt sich in diesem Fall unter Verwendung von (3.27), (3.29) und $e_{\min} = -e_{\max}$ als

$$ZB_{g_1} \approx 6,02(2^{t_{\exp}+1} - 1) \text{ dB} \quad (3.32)$$

und für die Festkommadarstellung als

$$ZB_f \approx 6,02 t \text{ dB} \quad (3.33)$$

wobei t die Festkommawortlänge (ohne Vorzeichen-Bit) ist. Verwendet man (3.27) und (3.30), so ergibt sich der Zahlenbereich durch

$$ZB_{g_2} \approx 6,02(t_{\max} + 2^{t_{\exp}+1} - 2) \text{ dB} \quad (3.34)$$

Als ein Beispiel nehme man an, daß $t=22$, $t_{\max}=15$ und $t_{\exp}=6$ Bits je-

weils ohne Verzeichen-Bit ist. Wir erhalten $ZB_{g1} \approx 764$ dB, $ZB_{g2} \approx 849,8$ dB und $ZB_f \approx 132$ dB. Bild 3.9 veranschaulicht die Ausdrücke (3.32), (3.33) und (3.34).

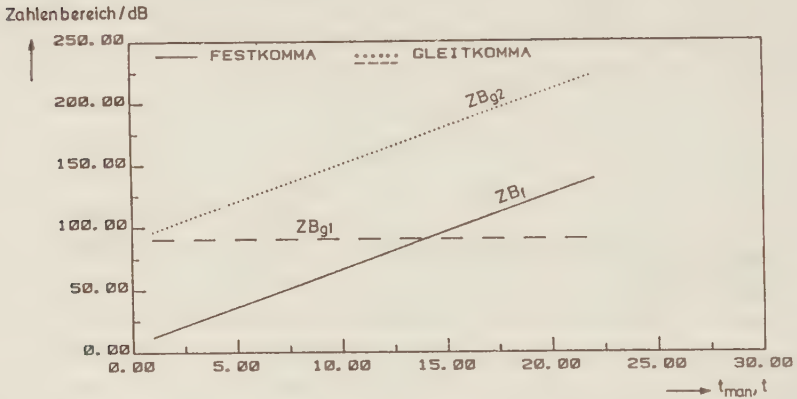


Bild 3.9a: Zahlenbereiche für Festkomma- und Gleitkommadarstellung ($t_{\text{exp}} = 3$ Bit ohne Vorzeichenbit).

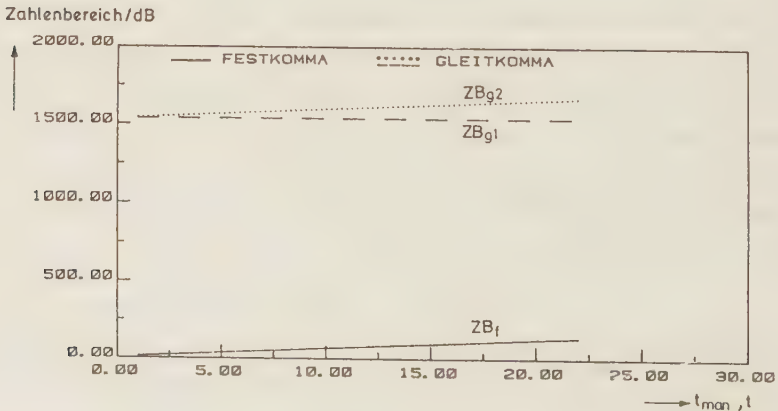


Bild 3.9b: Zahlenbereich für Festkomma- und Gleitkommadarstellung ($t_{\text{exp}} = 7$ Bit ohne Vorzeichenbit).

Wie man aus (3.32) und (3.33) und aus dem Bild 3.9c ersehen kann,

hängt der Zahlenbereich bei Gleitkommadarstellung exponentiell von der Exponentenwortlänge t_{exp} ab, wohingegen er bei der Festkommadarstellung linear mit der Wortlänge t ansteigt.

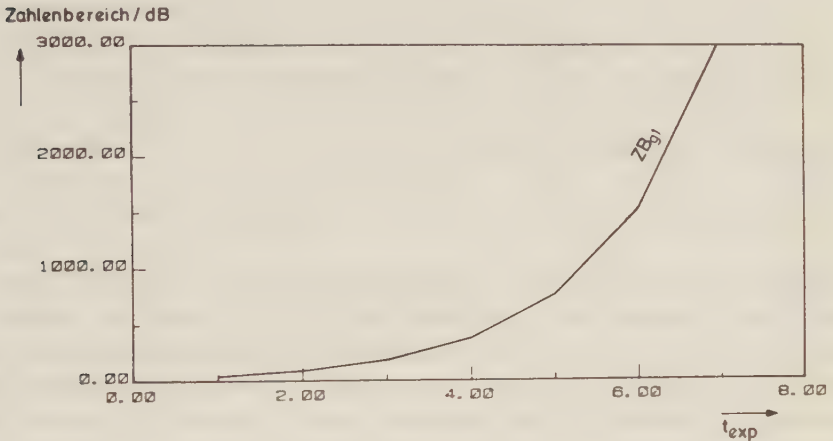


Bild 3.9c: Zahlenbereich für Gleitkommadarstellung nach (3.32).

Darüber hinaus ist die Genauigkeit einer Gleitkommadarstellung mit der Mantissenlänge verbunden. Das bedeutet, je größer die Wortlänge der Mantisse t_{man} ist, um so höher wird die Genauigkeit auch bei kleineren Signalen, da der relative Fehler ε_r (Tabelle 3.2) immer klein ist. Bei einer Festkommadarstellung bleibt der relative Fehler für kleinere Signale groß und demzufolge wird die Genauigkeit bei solchen Signalen geringer. Aus diesem Grund resultiert bei der Gleitkommadarstellung ein überlegenes Fehlerverhalten, besonders wenn t_{man} und t gleich sind.

Wir können nach der vorigen Diskussion davon ausgehen, daß man beim Gleitkommasystem durch eine geeignete Wahl der Wortlänge für jeweils die Mantisse und den Exponenten die Möglichkeit hat, sowohl die gewünschte Genauigkeit als auch den gewünschten Zahlenbereich zu erzielen.

Bei Digitalfiltern, in denen die Signalparameter in Festkommaarithmetik dargestellt sind, werden die Signale im Verlauf der Rechnung so skaliert, daß der verfügbare Zahlenbereich besser ausgenutzt werden kann. Man verwendet als Skalierungsfaktoren zweckmäßigerweise Potenzen von 2, nicht nur aus Aufwandsgründen sondern auch

aus Realisierungsgründen. Um dies zu klären nehme man an, daß die Zahl n und ihre reziproke Zahl $1/n$ als Dualzahlen mit endlicher Wortlänge darstellbar sind. Das bedeutet, n und $1/n$ lassen sich durch

$$n = m_1 \cdot 2^{e_1} \quad \text{bzw.} \quad \frac{1}{n} = m_2 \cdot 2^{e_2}$$

mit $m_i = a_i / 2^{b_i}$, a_i, b_i, e_i ganze Zahlen $i=1,2$

ausdrücken. Daraus folgt aber

$$m_1 \cdot m_2 = 2^{-(e_1+e_2)}, \quad \text{also} \quad a_1 \cdot a_2 = 2^{(b_1+b_2-e_1-e_2)}$$

Folglich sind a_1 und a_2 und damit auch n und $1/n$ Potenzen von 2 in der Form 2^e , wobei e eine ganze Zahl ist. In dieser Situation ist die Verwandtschaft zur Gleitkommadarstellung dann offensichtlich.

Im folgenden wollen wir die elementaren arithmetischen Gleitkommaoperationen, die bei der Rechnung im Digitalfilter benötigt werden, nämlich die Addition/Subtraktion und die Multiplikation, diskutieren.

3.8 Die elementaren arithmetischen Operationen

3.8.1 Addition und Subtraktion von zwei Gleitkommazahlen

Für die Addition oder Subtraktion von zwei Gleitkommazahlen

$x_1 = m_1 2^{e_1}$ und $x_2 = m_2 2^{e_2}$ gilt die folgende Beziehung

$$x_3 = x_1 \pm x_2 = m_1 2^{e_1} \pm m_2 2^{e_2} =$$

$$\left[m_1 2^{-\max(0, e_2 - e_1)} \pm m_2 2^{-\max(0, e_1 - e_2)} \right] 2^{\max(e_1, e_2)} \quad (3.35)$$

Aus der Beziehung (3.35) kann man ersehen, daß im allgemeinen eine der Mantissen vor der eigentlichen Addition oder Subtraktion mit einer entsprechenden Zweierpotenz multipliziert werden muß. Das bedeutet, um die Addition oder Subtraktion zweier Gleitkommazahlen durchführen zu können, muß die Mantisse der kleinsten Zahl um $|e_1 - e_2|$ Binärstellen geschoben werden, um die beiden Operandenexponenten anzugleichen. In diesem Fall sind die Exponenten der beiden Gleitkommazahlen x_1 und x_2 gleich. Dann können die beiden Mantissen addiert oder subtrahiert werden. Dazu braucht man einen Akkumulator

der Länge von $t_{\text{man}} + |e_1 - e_2|$ Bits und noch ein Bit extra, falls die neue resultierende Mantisse größer Eins ist. In diesem Fall, wenn Mantissenüberlauf auftritt, muß die Mantisse durch 2 geteilt und der Exponent des Ergebnisses, der sich aus dem Größenvergleich, d.h. aus dem größeren Exponenten ermitteln läßt, um Eins erhöht werden, damit die Mantisse die Bedingung (3.9) erfüllt. Bei der Durchführung der Subtraktion entsteht in der Regel ein nicht normiertes Ergebnis, d.h., die führende Eins in der resultierenden Mantisse steht nicht an der vorgeschriebenen Stelle. Zur Normierung des Ergebnisses wird die höchste Eins in der Mantisse lokalisiert. Dann wird die Mantisse nach links geschoben und der Exponent entsprechend erniedrigt.

Der wesentliche Mehraufwand eines Gleitkommaaddierers gegenüber einem Festkommaaddierer besteht darin, daß die Mantisse der kleinsten Zahl mit einem Zweierpotenzmultiplikator, dessen Exponent sich aus der Differenz der Operandenexponenten ergibt, multipliziert werden muß und daß eine längere Akkumulatorlänge zu gebrauchen ist. Ein integrierter Gleitkommaaddierer für eine Mantissenlänge von 16 Bits und eine Exponentenlänge von 6 Bits ist der TDC 1022 von TRW [36]. Die Rechenzeit liegt unter 100 ns (16 MHz). Ein Addierer, der im IEEE 754-Gleitkomma-Standard mit 32 Bits (24 Bits Mantisse und 8 Bits Exponent) arbeitet, ist der WTL 1032/1033 von Weitek [37]. Er kann 5 Millionen Operationen je Sekunde durchführen.

3.8.2 Multiplikation von zwei Gleitkommazahlen

Die nachstehende Beziehung gilt für die Multiplikation von zwei Gleitkommazahlen x_1 und x_2 :

$$x_4 = x_1 \cdot x_2 = m_1 \cdot m_2 \cdot 2^{e_1 + e_2} \quad (3.36)$$

Falls die neue Mantisse $m_1 \cdot m_2$ kleiner $1/2$ ist, muß sie mit 2 multipliziert und der Exponent um Eins erniedrigt werden, damit die Mantisse die Bedingung (3.9) erfüllt. Geht man davon aus, daß die beiden Mantissen m_1 und m_2 dieselbe Wortlänge t_{man} besitzen, ist die benötigte Akkumulatorlänge im allgemeinen $2t_{\text{man}}$ um die nach der Multiplikation resultierende Mantisse darstellen zu können. Der Exponent der Multiplikation ergibt sich aus der Summe der Operandenexponenten.

Im Vergleich zu dem Festkomma-multiplizierer muß beim Gleitkomma-multiplizierer zusätzlich die Summe der Operandenexponenten gebildet werden. Auch ein Normierungsschnitt kann erforderlich sein. Ein Gleitkomma-Multiplizierer, der 5 Millionen Operationen je Sekunde im 32 Bit IEEE-Format durchführt, ist der WTL 1032/1033 von Weitek. Ein besonders hohes Maß an Flexibilität bietet der Signalprozessor UPD 77230 von NEC. Er besitzt die Wortlänge von 47 oder 24 Bits für die Mantisse und 8 Bits für den Exponenten jeweils mit Vorzeichenbit. Die Rechenzeit für die Multiplikation und Akkumulation ist 150 ns.

3.9 Komplexität der Hardwarerealisierung einer Gleitkommaarithmetik

Im Unterabschnitt 3.7 wurde erwähnt, daß bei einer Gleitkomma-zahlendarstellung der Zahlenbereich groß und die Genauigkeit auch bei kleineren Signalen hoch ist. Um dies zu erzielen, muß man den Hardwarerealisierungsaufwand in Kauf nehmen. Im folgenden wird eine Schätzung der Hardwarekomplexität eines Multiplizierers und eines Addierers sowohl im Fall der Festkomma-zahlendarstellung als auch bei Gleitkommaarithmetik gegeben, unter der Annahme, daß der Hardwareaufwand des Speichers nicht in Betracht gezogen wird. Der Grund besteht darin, daß die benötigte Speicherlänge in beiden Fällen der Darstellung, um ein gewünschtes Systemverhalten zu erzielen, ungefähr dieselbe bleibt. Darüber hinaus läßt sich der Hardwarerealisierungsaufwand eines Multiplizierers oder eines Addierers in Abhängigkeit der Gatterzahl (=Anzahl der benötigten Gatter) ausdrücken [41].

Bei einem Festkommaaddierer besitzen die Operandensignale im allgemeinen die Wortlänge von t Bits plus ein Bit für das Vorzeichen. Unter der Annahme, daß der Addierer aus Volladdiererzellen besteht, läßt sich der Realisierungsaufwand A_f durch,

$$A_f = 10 (t+1) \quad (3.37)$$

schätzen. Im Gegensatz zur Festkommaaddition muß man, wie schon erwähnt, bei der Gleitkommaaddition zunächst die beiden Exponenten angleichen, bis die beiden zu addierenden Gleitkommazahlen den größeren Exponenten besitzen. Dazu bildet man die Differenz der beiden Exponenten und verschiebt die Mantisse, die zu dem kleineren Exponenten gehört, um entsprechend viele Bits nach rechts. Diese Schie-

beoperation wird üblicherweise mit einem Schieberegister durchgeführt. Dabei ergibt sich pro Takt eine Verschiebung um eine Stelle. Nachteilig ist, daß man eine Ablaufsteuerung benötigt, um die Verschiebung um eine vorwählbare Stellenzahl vorzunehmen. Dieselbe Schiebeoperation läßt sich ohne getaktete Ablaufsteuerung durchführen, indem man eine entsprechende Schaltung mit Multiplexern aufbaut. Diese Schaltung wird "Barrel shifter" genannt.

Möchte man eine höhere Arbeitsgeschwindigkeit, etwa wie bei Festkommaaddition, erreichen, so muß die Schiebeoperation mit einem "Barrel shifter" durchgeführt werden.

Nach der Subtraktion entsteht im allgemeinen ein nicht normiertes Ergebnis. Zur Normierung wird die Mantisse nach links geschoben und der Exponent entsprechend verkleinert.

Der minimale Realisierungsaufwand eines Gleitkommaaddierers läßt sich durch [41]

$$A_g = t_{\text{man}}^2 / 2 + 24.5 t_{\text{man}} + 345 \quad (3.38)$$

schätzen, unter der Annahme, daß $t_{\text{exp}} = 7$ Bits ist.

Im Fall eines Festkommamultiplizierers, der aus einer Anordnung von Volladdiererzellen besteht, läßt sich der Hardwarerealisierungsaufwand M_f folgendermaßen darstellen:

$$M_f = 10(t_1-1)t_2 + t_1 t_2, \quad (3.39)$$

wobei t_1 die Wortlänge eines Operanden und t_2 die Wortlänge des zweiten Operanden ist. (Unterschiedliche Wortlängen für die Operanden sind relevant in Digitalfiltern, da die Koeffizienten, insbesondere im Fall der Wellendigitalfilter, in kürzerer Wortlänge als die Signalwortlänge dargestellt werden). Gleitkommamultiplizierer benötigen außer der Multiplikation der beiden Operandenmantissen zusätzliche Hardware für die Addition der Operandenexponenten und für die Normierung. Der Hardwarerealisierungsaufwand M_g eines Gleitkommamultiplizierers läßt sich durch folgende Beziehung ausdrücken:

$$M_g = 10(t_{\text{man1}}-1)t_{\text{man2}} + t_{\text{man1}}t_{\text{man2}} + t_{\text{man1}} + t_{\text{man2}} + 10(t_{\text{exp}}+1) \quad (3.40)$$

Damit ist t_{man1} die Mantissenwortlänge des ersten Operanden und t_{man2} die Mantissenwortlänge des zweiten Operanden. In Bild 3.10 und 3.11 ist der Zusammenhang zwischen dem Realisierungsaufwand sowohl eines Addierers als auch eines Multiplizierers und der Wortlänge $t(t_{\text{man}})$ eingetragen. Es bleibt anzumerken, daß

der Realisierungsaufwand eines Gleitkommaaddierers höher ist als der Aufwand eines Festkommaaddierers, wohingegen der Aufwand beim Gleitkomma- und Festkommamultiplizierer für $t=t_{\text{man}}$ etwa derselbe bleibt. Das bedeutet, daß die Exponenten vergleichsweise wenig Einfluß auf den gesamten Realisierungsaufwand eines Multiplizierers haben.

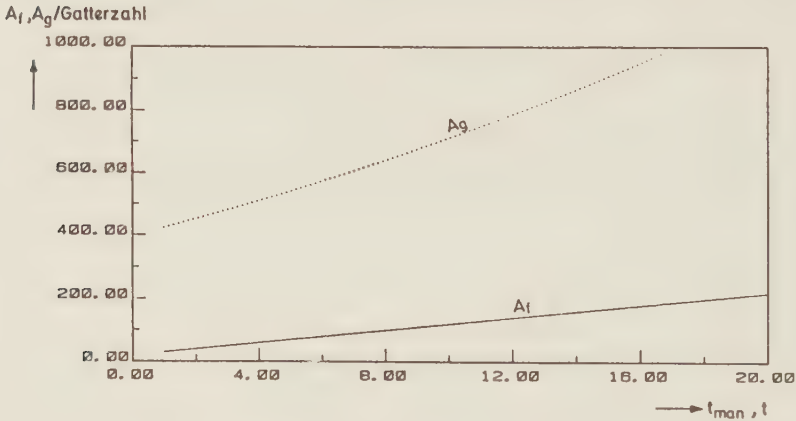


Bild 3.10: Realisierungsaufwand eines Festkomma- und Gleitkommaaddierers in Abhängigkeit der Wortlänge ($t_{\text{exp}}=7\text{Bits}$).

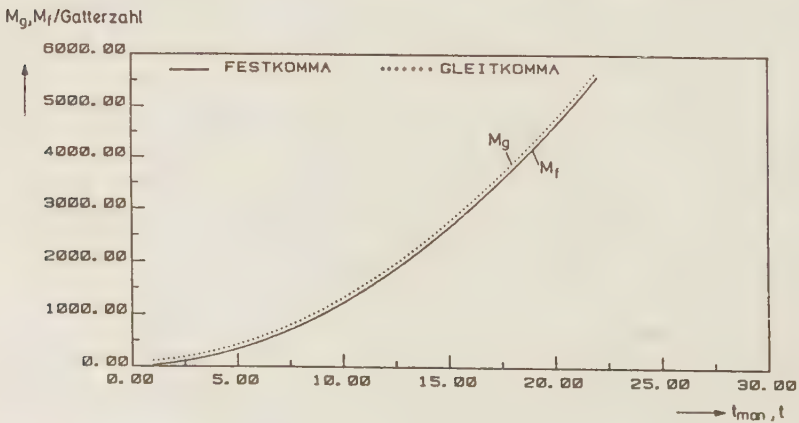


Bild 3.11: Realisierungsaufwand eines Festkomma- und Gleitkommamultiplizierers in Abhängigkeit der Wortlänge ($t_{\text{exp}}=7\text{Bits}$).

Es sei an dieser Stelle ausdrücklich darauf hingewiesen, daß der gesamte Realisierungsaufwand eines Systems, z.B. eines Digitalfilters, in der Bedeutung über den bloßen Aufwand eines einzelnen Addierers und Multiplizierers noch weit hinausgeht.

4. Passivitätseigenschaften von Gleitkomma-Wellendigitalfiltern

4.1 Vorbemerkungen

In rekursiven Digitalfiltern, zu denen Wellendigitalfilter gehören, treten als Folge der unvermeidlichen Endlichkeit der Signalwortlänge Fehler auf, die nicht linear sind. Diese entstandenen Nichtlinearitäten lassen sich in zwei Kategorien klassifizieren. Die Nichtlinearitäten der ersteren Kategorie werden durch die notwendigen Rundungs- und/oder Schneideoperationen des Ergebnisses nach bestimmten arithmetischen Operationen hervorgerufen. In diesem Zusammenhang wird von der sogenannten Granularität gesprochen. Die der zweiten werden dadurch verursacht, daß das Ergebnis einer arithmetischen Operation den verfügbaren Zahlenbereich überschreitet (Überlauf).

Bekanntlich können bei abgeschaltetem Eingangssignal parasitäre Schwingungen in einem rekursiven Digitalfilter entstehen [42-46]. Diese Schwingungen werden auch als Grenzzyklusschwingungen bezeichnet und lassen sich je nach ihrer Entstehung in Granular- und Überlaufschwingungen unterteilen. Die Schwingungen der ersten Art werden als Feinschwingungen bezeichnet. Die der zweiten werden Grobschwingungen genannt [13].

Die Granularschwingungen betreffen die letzten Bits der Signalewörter und haben im Fall der Festkommarealisierung eines Filters Amplituden, die weit unterhalb der Überlaufgrenze bleiben. Jedoch sind diese Schwingungen störend, insbesondere wenn sie nur wenige Frequenzkomponenten umfassen. In diesem Fall muß die Unterdrückung dieser Schwingungen in Betracht gezogen werden. Die Überlaufschwingungen zerstören völlig die Betriebsfähigkeit eines Digitalfilters, da ihre Amplituden in der Größenordnung der maximal darstellbaren Signale liegen. Deshalb müssen sie unter allen Umständen verhindert werden.

In rekursiven Festkomma-Digitalfiltern üblicher Art, die als Kaskadenschaltung von Gliedern erster und zweiter Ordnung aufgebaut werden, lassen sich die Überlaufschwingungen durch Benutzung der Sättigungs- oder Dreieckskennlinien an geeigneten Stellen (an den Multiplizierereingängen) in solchen Gliedern zur Überlaufkorrektur vermeiden. Die Granularschwingungen sind besonders kritisch, wenn das Runden verwendet wird. Die Verwendung von Betragsschneiden zur Unterdrückung dieser Schwingungsart ist auch nur begrenzt wirksam.

Im Gegensatz dazu lassen sich Wellendigitalfilter, die mit Festkommaarithmetik realisiert sind, als Folge ihrer Pseudopassivitätseigenschaften sehr einfach völlig frei von parasitären Schwingungen realisieren [26,27,49]. In [26] lassen sich hinreichende jedoch nicht notwendige Bedingungen herleiten, um die Stabilität eines Wellendigitalfilters nach dem Abschalten des Eingangssignals stets sicherzustellen. Dies bedeutet, daß eine teilweise Verletzung der Bedingungen keineswegs Instabilität verursachen muß. Die Stabilität der Wellendigitalfilter kann sogar gewährleistet werden, wenn das Filter innerhalb einer Schleife arbeitet [50]. Diese interessante Eigenschaft von Wellendigitalfiltern folgt daraus, daß sich diese Filter durch technisch einfach zu realisierende Maßnahmen wie passive Schaltungen verhalten.

Bei normalem Betrieb, d.h., bei eingeschaltetem Eingangssignal, kann durch die entstehenden Überläufe das gewünschte Filterverhalten beeinträchtigt werden. Eine der wichtigsten Auswirkungen der Überläufe unter normalem Betrieb besteht darin, daß die Antwortstabilität des Filters nicht gewährleistet werden kann [28]. Unter Antwortstabilität versteht man, daß die durch einen Überlauf verursachten Störungen des Ausgangssignals auch dann mit der Zeit abklingen müssen, wenn das Eingangssignal nicht auf einen kleineren Pegel abgesenkt wird. Obwohl bisher nicht bekannt ist, wie die Antwortstabilität für die üblichen rekursiven Digitalfilter gewährleistet werden kann, lassen sich Wellendigitalfilter als Folge ihrer inkrementalen Pseudopassivität [30] durch Verwendung geeigneter Überlaufkennlinien (unter Vernachlässigung der Granularität) an den Adaptorausgängen vollständig antwortstabil realisieren [30].

Bei der Digitalfilterrealisierung mit Gleitkommaarithmetik, in der die Signalparameter als Gleitkommazahlen dargestellt werden, können parasitäre Schwingungen sogar mit großen Amplituden entstehen [47]. Die Analyse dieses Problems findet sich in [38,47,48]. Die entsprechende Theorie wurde für die rekursiven Digitalfilterglieder erster und zweiter Ordnung durchgeführt, da ein rekursives Digitalfilter als Kaskadenschaltung von solchen Gliedern aufgebaut werden kann.

In diesem Kapitel werden wir die Pseudopassivitätseigenschaften der verschiedenen Wellendigitalfilterelemente diskutieren, um zu zeigen, wie im Fall einer Zweierkomplement-Gleitkommaarithmetik Wellendigitalfilter völlig frei von parasitären Schwingungen und

antwortstabil realisiert werden können. Bevor wir auf die Analyse eingehen, wollen wir einige grundlegende Definitionen kurz zusammenfassen.

4.2 Stabilitätskriterium

Wir betrachten das in Bild 4.1 gezeichnete n-Tor-Wellendigitalfilter. Dabei enthält N nur die zur Berechnung der Ausgangssignale b_v ($v=1,2,\dots,n$) benötigten arithmetischen Operationen. Die Pseudoleistung $P_v(t_m)$, die durch das Tor v zum Zeitpunkt

$$t_m = t_0 + mT, \quad m = \dots, -2, -1, 0, 1, 2, \dots$$

eintritt, läßt sich durch

$$P_v(t_m) = [a_v^2(t_m) - b_v^2(t_m)] G_v \quad (4.1)$$

definieren [25], wobei a_v das Eingangssignal, b_v das Ausgangssignal oder die Antwort und G_v den Leitwert des Tors v bezeichnet. Außerdem ist t_0 ein beliebig wählbarer Bezugszeitpunkt (meist $0 < t_0 < T$ [24]). Die entsprechende Pseudoenergie wird durch

$$W_v(t_m) = \sum_{\mu=0}^m P_v(t_\mu), \quad m > 0 \quad (4.2)$$

gegeben. Die gesamte Pseudoleistung $P(t_m)$, die von dem n-Tor N aufgenommen wird, läßt sich ausdrücken als

$$P(t_m) = \sum_{v=1}^n [a_v^2(t_m) - b_v^2(t_m)] G_v. \quad (4.3)$$

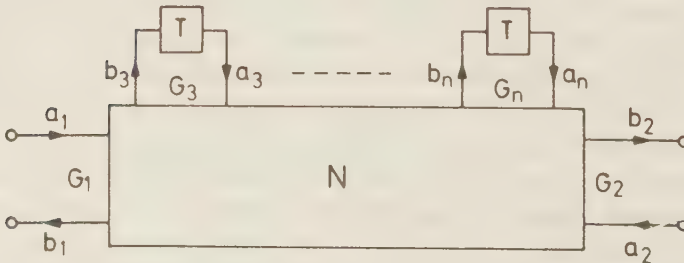


Bild 4.1: Ein n-Tor-Wellendigitalfilter.

Entsprechend gilt für die gesamte Pseudoenergie $W(t_m)$

$$W(t_m) = \sum_{v=1}^m W_v(t_m) = \sum_{v=1}^n \sum_{\mu=0}^m P_v(t_\mu) \quad . \quad (4.4)$$

Gilt die Bedingung

$$P(t_m) \geq 0 \quad \forall a_v(t_m) \quad ,$$

so wird das n-Tor N als pseudopassiv bezeichnet [25]. Falls das Gleichheitszeichen für alle $a_v(t_m)$ gilt, spricht man von Pseudoenergieneutralität des n-Tores N. Es sei bemerkt, daß diese Definitionen auch gelten, wenn das n-Tor N nichtlinear ist. Um die Pseudopassivität eines Wellendigitalfilters unter nichtlinearen Bedingungen zu gewährleisten, ist in [26] eine Bedingung für die Ausgangssignale b_v gegeben:

$$|b_v| \leq |b_{v0}| \quad . \quad (4.5)$$

Dabei sind b_{v0} die Ausgangssignale des n-Tores N unter idealen linearen Bedingungen, d.h., alle arithmetischen Operationen innerhalb des n-Tores N müssen exakt durchgeführt werden, und b_v sind die berechneten Ausgangssignale unter nichtlinearen Bedingungen, d.h. mit begrenzter Wortlänge. a_v müssen in den beiden Fällen dieselben bleiben. Außerdem ist hierbei angenommen, daß N unter linearen Bedingungen pseudopassiv ist.

In der Praxis besteht das n-Tor N im allgemeinen aus Adaptoren, idealen Übertragern, Gyrotoren und Zirkulatoren. Es ist bekannt, daß sich die übertragenen Pseudoleistungen durch zwei miteinander verbundene Tore zu Null ergänzen [25]. Daraus folgt, daß die gesamte von dem n-Tor-Wellendigitalfilter aufgenommene Pseudoleistung gleich der Summe der von den einzelnen Elementen aufgenommenen Pseudoleistungen ist [26]. Aus diesem Grunde ist ein Wellendigitalfilter dann pseudopassiv, wenn seine Bauelemente individuell pseudopassiv sind. Man spricht in diesem Fall von konkreter Pseudopassivität. Aufgrund dieser Tatsache geht man davon aus, daß die Pseudopassivität eines Wellendigitalfilters sichergestellt werden kann, wenn für jedes Wellendigitalbauelement die Bedingung (4.5) eingehalten wird.

Von großer Bedeutung ist die sogenannte inkrementale Pseudopassivität [29]. Sie führt zu einer Stetigkeit des Übertragungsverhaltens eines Filters unter nichtlinearen Bedingungen. Das bedeutet, daß kleine Änderungen des Eingangssignals nur

kleine Änderungen des Ausgangssignals zur Folge haben können, und allgemeiner, daß geringfügige Übersteuerung am Eingang nur geringfügige Verzerrung am Ausgang verursachen kann. Um dies zu erklären, nehmen wir an, daß zwei voneinander unabhängige Eingangssignale a_u und a'_u auf N wirken. Damit ergeben sich die entsprechenden Ausgangssignale b_u und b'_u . Die inkrementale Pseudoleistung P_Δ ist durch

$$P_\Delta(t_m) = \sum_{u=1}^n [[a_u(t_m) - a'_u(t_m)]^2 - [b_u(t_m) - b'_u(t_m)]^2] G_u \quad (4.6)$$

definiert. Für die inkrementale Pseudopassivität gilt

$$P_\Delta(t_m) \geq 0 \quad \text{für alle } a_u(t_m) \text{ und } a'_u(t_m).$$

Um die inkrementale Pseudopassivität für ein Eintor sicherzustellen, muß somit

$$|a_u(t_m) - a'_u(t_m)| \geq |b_u(t_m) - b'_u(t_m)| \quad (4.7)$$

gelten. Beziehung (4.7) kann so interpretiert werden, daß sich die zwei Ausgangssignale b_u und b'_u wenig voneinander unterscheiden, solange sich die zugehörigen Eingangssignale nur wenig voneinander unterscheiden.

Darüber hinaus ist die inkrementale Pseudopassivität eine hinreichende Bedingung für die Antwortstabilität eines Wellendigitalfilters. Um dies näher zu veranschaulichen, nehmen wir an, daß bei anliegendem Eingangssignal ein einmaliger Überlauf auftritt. Das bedeutet, daß das Eingangssignal einen überlauffreien stationären Betrieb zuläßt. Nun seien $a'_u(t_m)$ ($u=1,2$) in (4.6) die Eingangssignale, die die Überläufe beim Zeitpunkt t_m verursachen. Das 2-Tor aus Bild 4.1 wird als antwortstabil bezeichnet, falls

$$\lim_{m \rightarrow \infty} [b_u(t_m) - b'_u(t_m)] = 0, \quad u=1,2 \quad (4.8)$$

für alle $a_u(t_m)$ und alle möglichen Anfangszustände auch gilt. Tatsächlich sind Wellendigitalfilter durch Verwendung geeigneter Überlaufkennlinien an den Adaptoraustritten inkremental pseudopassiv.

Wie man aus einer Verallgemeinerung der Beziehung (4.7) entnehmen kann, ist die inkrementale Pseudopassivität eine hinreichende Bedingung für die Antwortstabilität. Wählt man nämlich $a_u(t_m) = a'_u(t_m)$, so konvergieren $b_u(t_m)$ und $b'_u(t_m)$, $u=1,2$, gegeneinander. Folglich wird die Bedingung (4.8) erfüllt [30].

4.3 Pseudopassivitätseigenschaften der Bauelemente von WDF

4.3.1 Vorbemerkungen

Wie schon zu Beginn dieses Kapitels erwähnt wurde, führt die Begrenzung der Signalwortlänge zu einem nichtlinearen Verhalten der Wellendigitalbauelemente. In diesem Abschnitt wollen wir untersuchen, welchen Einfluß die Begrenzung der Signalwortlänge auf die Pseudopassivitätseigenschaften der Elemente im Fall der Gleitkommaarithmetik hat und wie man die Pseudopassivität garantieren kann.

Es wurde gezeigt [25], daß die Bauelemente von Wellendigitalfiltern unter der Annahme einer unendlichen Koeffizienten- und Signalwortlänge pseudopassiv sind. Im folgenden nehmen wir an, daß die Begrenzung der Koeffizientenwortlänge durch diskrete Optimierung so vorgenommen werden, daß die Pseudopassivität der Bauelemente nicht beeinträchtigt wird.

Da die Wellenflußdiagramme des Widerstands, des Zirkulators, des Leerlaufs, des quasi-reziproken Leitungselements und der Kapazität keine arithmetischen Operationen enthalten, behalten diese Elemente auch unter Berücksichtigung einer endlichen Signalwortlänge sowohl ihre Pseudopassivität als auch ihre inkrementale Pseudopassivität. Dieselbe Aussage trifft offenbar auch für den Gyrator, den Kurzschluß und die Induktivität unter der Voraussetzung zu, daß die Multiplikation mit -1 fehlerfrei ausgeführt werden kann. In den folgenden Unterabschnitten betrachten wir die Pseudopassivität der Bauelemente, deren Signalflußdiagramme arithmetische Operationen beinhalten.

4.3.2 Pseudopassivitätseigenschaften der Adaptoren

In Abschnitt 4.2 wurde erwähnt, daß ein Wellendigitalfilter unter Berücksichtigung einer endlichen Signalwortlänge pseudopassiv bleibt, solange die Bauelemente individuell ihre Pseudopassivitätseigenschaften beibehalten. Im Fall eines Adaptors ist die Situation aber weniger günstig. Der Grund besteht darin, daß innerhalb eines Adaptors mehrere arithmetische Operationen durchzuführen sind. Besonders nachteilig wirken sich mehrere hintereinander angeordnete Additionen und Multiplikationen aus. Daher ist es sinnvoll, einen Adaptor mit vielen arithmetischen Operationen soweit wie möglich in weniger komplexe Adaptoren zu zerlegen. Es wurde gezeigt,

daß jeder Parallel- und Serienadaptor durch die sogenannten elementaren Adaptoren realisiert werden kann [51]. Die elementaren Adaptoren sind die 3-Tor-Adaptoren mit einem reflexionsfreien Tor sowie der 2-Tor-Adaptor.

Es wurde gezeigt, daß die Pseudopassivitätseigenschaft eines n-Tor-Adaptors im Fall der Gleitkommaarithmetik eingehalten werden kann [52]. Dafür muß man jeden Adaptoraustritt zweimal berechnen. Das bedeutet, daß in der Tat mehrere Additionen und Multiplikationen für z.B. 3-Tor-Adaptor mit einem reflexionsfreien Tor durchzuführen sind. Außerdem können die Quantisierungsfehler den Dynamikbereich des gesamten Filters beeinträchtigen. Wir wollen zeigen, daß die Pseudopassivitätseigenschaften eines 2- und 3-Tor-Adaptors im Fall der Gleitkommarealisierung mit endlicher Signalwortlänge gewährleistet werden können, ohne die Adaptoraustritte zweimal zu berechnen.

Für die Gleitkommarealisierung eines Adaptors muß die Anzahl der Addierer möglichst klein sein, da, wie schon erwähnt wurde, die Durchführung einer exakten Addition/Subtraktion einen langen Akkumulator benötigt (im Gegensatz zur Festkommaaddition). (Als Beispiel betrachten wir zwei Gleitkommazahlen $x_1 = m_1 \cdot 2^{15}$ und $x_2 = m_2 \cdot 2^{-15}$, wobei m_1 und m_2 jeweils die Wortlänge von 12 Bits (ohne Vorzeichenbit) besitzen. Will man die beiden Zahlen x_1 und x_2 exakt addieren, so wird ein Akkumulator der Länge von 42 Bits (ohne Vorzeichenbit) benötigt. Aus diesem Grunde ist die Realisierung der Adaptoren direkt durch ihre Streumatrizen günstiger für die Gleitkommaarithmetik. Im folgenden wollen wir diese Tatsache erörtern.

4.3.2.1 Pseudopassivität des 2-Tor-Adaptors

Der Beziehung (2.15) kann man die folgenden Gleichungen für die Adaptoraustritte entnehmen:

$$\begin{aligned} b_1 &= -\gamma a_1 + \gamma_2 a_2 \\ b_2 &= \gamma_1 a_1 + \gamma a_2 \end{aligned} \quad , \quad (4.9)$$

mit
$$\gamma = \frac{R_1 - R_2}{R + R} = 1 - \gamma_1 = \gamma_2 - 1$$

Bild 4.2 zeigt das zur Realisierung (4.9) gehörende Signalflußdiagramm.

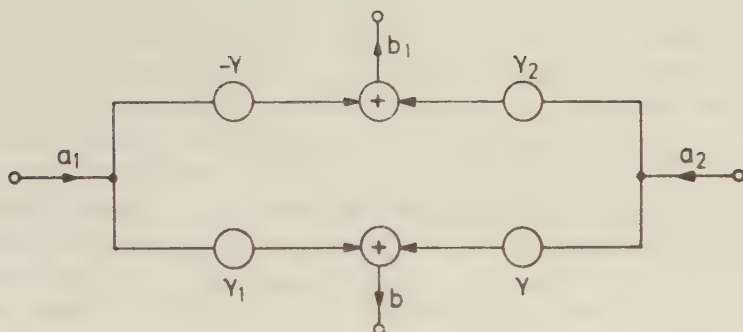


Bild 4.2: Das zur Realisierung von (4.9) gehörende Signalflußdiagramm.

Dieselbe Aussage gilt auch für einen 2-Tor-Serienadaptor, der direkt durch seine Streumatrix realisiert wird.

Es sei erinnert, daß die Multipliziererkoeffizienten in einem Wellendigitalfilter mit einfacher kurzer Wortlänge in Festkommazahlendarstellung realisiert werden können [13-20]. Da diese Koeffizienten nicht größer als 2 sind, können wir ohne Beschränkung der Allgemeinheit annehmen, daß Y , Y_1 und Y_2 Gleitkommazahlen aus dem Gleitkommasystem

$$GK = GK(2, t'_{\text{man}}, e'_{\text{min}}, 1) \quad (4.10a)$$

sind. Ferner seien die Signale a_1, a_2, b_1 und b_2 aus dem Gleitkommasystem

$$GS = GS(2, t_{\text{man}}, e_{\text{min}}, e_{\text{max}}) \quad (4.10b)$$

Wir gehen im folgenden davon aus, daß die Multiplikation in dem zu betrachtenden Adaptor derart ausgeführt wird, daß das Ergebnis nach jedem Multiplikationsschritt ohne Quantisierungsfehler dargestellt wird. Das bedeutet, die resultierenden Ergebnisse $Y a_1, Y a_2, Y_1 a_1, Y_2 a_2$ sind im allgemeinen aus dem Gleitkommasystem

$$GS = GS(2, t_{\text{man}} + t'_{\text{man}}, e_{\text{min}} + e'_{\text{min}}, e_{\text{max}} + 1) \quad (4.11)$$

Will man die Pseudopassivität des Adaptors einhalten, so muß man zwei Gleitkommazahlen aus dem in (4.11) gegebenen Gleitkommasystem so addieren, daß das Ergebnis betragsmäßig die Bedingung (4.5) erfüllt. Um dies zu verwirklichen kann man mit einem Akkumulator der

Länge $l = t_{\text{man}} + t'_{\text{man}} + e_{\text{max}} + 1 + e_{\text{min}} + e'_{\text{min}}$ Bits (ohne Vorzeichenbit) zusätzlich einem extra Bit als Schutzbit gegen Mantissenüberlauf die Gleitkommaaddition exakt durchführen [53] und dann den Betrag des Endergebnisses auf t_{man} zurückschneiden. Eine andere Möglichkeit besteht darin, daß ein Akkumulator verwendet wird, der den Bereich vom kleinsten bis zum größten vorkommenden Exponenten überdeckt. Die zu addierenden Gleitkommazahlen werden also in ein Festkommasystem ausreichender Länge abgebildet [53]. Für diesen Akkumulator wird eine Länge $l = t_{\text{man}} + t'_{\text{man}} + e_{\text{max}} + 2 + e_{\text{min}} + e'_{\text{min}}$ Bits benötigt. Es können also die beiden Summanden in den Akkumulator exakt aufaddiert werden.

Im folgenden wollen wir ohne Benutzung von langem Speicher versuchen, die Bedingung (4.5) zu erfüllen. Dazu betrachten wir die Berechnung des Adaptorausgangssignals b_1 . Für das Signal b_2 läuft die Berechnung auf ähnliche Weise. Das Ausgangssignal b_{10} läßt sich als eine Summe von zwei Gleitkommazahlen wie folgt

$$b_{10} = -\gamma a_1 + \gamma_2 a_2 = x_1 + x_2 \quad (4.12)$$

beschreiben, wobei $x_1 = m_1 \cdot 2^{e_1}$ und $x_2 = m_2 \cdot 2^{e_2}$ zwei Gleitkommazahlen aus dem in (4.11) gegebenen Zahlensystem GS sind. Ferner nehmen wir an, daß die beiden Zahlen x_1 und x_2 , d.h. m_1, e_1, m_2 und e_2 , im Zweierkomplementkode dargestellt sind. Unter der Annahme, daß die Akkumulatorlänge $t_{\text{man}} + t'_{\text{man}}$ Bits (ohne Vorzeichenbit) ist (plus 1 Bit als Schutzbit gegen Mantissenüberlauf), betrachten wir die Summe $x_1 + x_2$ in allen möglichen Fällen:

a) $x_1 + x_2 > 0$

Ohne Beschränkung der Allgemeinheit kann man annehmen, daß $e_1 \geq e_2$ ist. In diesem Fall muß die Mantisse m_2 um $(e_1 - e_2)$ Binärstellen geschoben werden, um die Addition von x_1 und x_2 ausführen zu können. Diese Schiebeoperation beeinflusst im allgemeinen den Betrag der Zahl x_2 , so daß x_2 in eine neue Zahl x'_2 übergeht. Diese neue Zahl läßt sich durch

$$x'_2 = x_2 (1 - \epsilon_0) = x_2 - \epsilon_{\text{abs}} \quad (4.13)$$

darstellen. Dabei ist ϵ_0 durch die folgenden Schranken

$$0 \leq \varepsilon_0 \leq 1 \quad \text{für } x_2 > 0 \quad (4.14a)$$

$$-2^{-t} < \varepsilon_0 \leq 0 \quad \text{für } x_2 < 0 \quad (4.14b)$$

gegeben, wobei $t = t_{\text{man}} + t'_{\text{man}}$ die Akkumulatorlänge bezeichnet. Es ist einfach zu prüfen, daß $\varepsilon_0 = 1$ ist, wenn $e_1 - e_2 \geq t_{\text{man}} + t'_{\text{man}}$ ist, oder anders ausgedrückt, wenn die Zahl x_2 wegen der erforderlichen Schiebeoperation aus dem Akkumulator herausfällt. Der in (4.13) gegebene absolute Fehler ε_{abs} läßt sich durch

$$0 \leq \varepsilon_{\text{abs}} < 2^{-t+e_1} \quad (4.15)$$

angeben. Das mit t Bits berechnete Ergebnis $(x_1 + x'_2)$ läßt sich mit Hilfe von (4.13) als

$$b_{11} = x_1 + x'_2 = x_1 + x_2 - \varepsilon_{\text{abs}} \quad (4.16)$$

schreiben, wobei $b_{11} = m_{11} \cdot 2^{e_1}$ ist. Wie man aus (4.16) ersehen kann, gilt die folgende Beziehung für die positive Summe

$$b_{11} = x_1 + x'_2 < x_1 + x_2 = b_{10} \quad (4.17)$$

Schneidet man die Mantisse von b_{11} auf t_{man} Bits zurück, so erhält man das Ausgangssignal b_1 , das mit Hilfe von (4.17) die folgende Bedingung erfüllt:

$$b_1 = [b_{11}]_q \leq b_{11} \leq b_{10} \quad (4.18)$$

also wegen $b_{11} > 0$ auch $|b_1| \leq |b_{10}|$.

b) $x_1 + x_2 < 0$

Unter derselben Annahme, daß $e_1 \geq e_2$ ist, gilt die Gleichung (4.16) auch in diesem Fall ohne Begrenzung der Allgemeinheit. Es ist aus (4.16) zu ersehen, daß die folgende Beziehung für die negative Summe gilt:

$$|b_{11}| = |x_1 + x'_2| = |x_1 + x_2 - \varepsilon_{\text{abs}}| \geq |x_1 + x_2| = |b_{10}|$$

da ε_{abs} stets positiv ist (vgl. (4.15)). Addiert man den maximalen absoluten Fehler 2^{-t+e_1} zu beiden Seiten der (4.16), so erhält man

$$b_{11+2}^{-t+e_1} = x_1 + x_2 + (2^{-t+e_1 - \epsilon_{\text{abs}}}) \quad . \quad (4.19)$$

Der Term 2^{-t+e_1} wird als Korrekturterm bezeichnet.

Aus (4.19) folgt für den Betrag der negativen Summe

$$|b_{11+2}^{-t+e_1}| < |x_1 + x_2| = |b_{10}| \quad , \quad (4.20a)$$

oder

$$|(m_{11+2}^{-t}) 2^{e_1}| < |b_{10}| \quad . \quad (4.20b)$$

Schneidet man den Betrag der resultierenden Mantisse (m_{11+2}^{-t}) auf t_{man} Binärstellen zurück, so läßt sich die folgende Bedingung erfüllen

$$|b_1| = |[b_{11+2}^{-t+e_1}]_q| < |b_{10}| \quad . \quad (4.21)$$

Es sei an dieser Stelle darauf hingewiesen, daß der Korrekturterm 2^{-t+e_1} nicht benötigt wird, wenn $e_1 = e_{\text{max}}$ ist und die Summe der beiden Mantissen $(m_1 + m_2)$ überläuft.

Aus der vorigen Analyse können wir ersehen, daß ein 2-Tor-Adaptor durch geeignete Maßnahmen bei der Berechnung seiner beiden Ausgänge pseudopassiv realisiert werden kann. Daraus folgt, daß die Pseudopassivität der Brücken-Wellendigitalfilter [54,55], die unter ausschließlicher Verwendung 2-Tor-Adaptoren aus symmetrischen Referenzfiltern in Brückenschaltung hergeleitet werden [58], erzielt werden kann. Diese Filterstrukturen sind, trotz ihrer hohen Empfindlichkeit im Sperrbereich gegen Koeffizientenänderung, von praktischem Interesse, da sie im allgemeinen eine geringere Addierer- und Multipliziererzahl als Abzweig-Wellendigitalfilter erfordern, um die gleiche Übertragungsfunktion zu realisieren [56]. Außerdem werden in der Praxis Abtastratenänderungen in diesen Strukturen stets einfach ermöglicht [57].

4.3.2.2 Pseudopassivität des 3-Tor-Adaptors

Wir beschränken uns auf 3-Tor-Adaptoren, da, wie schon erwähnt wurde, jeder n-Tor-Adaptor durch die sogenannten elementaren Adaptoren realisiert werden kann. Die Erhaltung der Pseudopassivität eines 2-Tor-Adaptors wurde im Unterabschnitt 4.3.2.1 diskutiert. Im

folgenden wollen wir untersuchen, inwieweit die 3-Tor-Adaptoren ihre Pseudopassivität unter Berücksichtigung der endlichen Signallängte erhalten können.

Für einen allgemeinen 3-Tor-Paralleladaptor gelten die folgenden Beziehungen

$$b_1 = (\gamma_1 - 1)a_1 + \gamma_2 a_2 + \gamma_3 a_3 \quad (4.22a)$$

$$b_2 = \gamma_1 a_1 + (\gamma_2 - 1)a_2 + \gamma_3 a_3 \quad (4.22b)$$

$$b_3 = \gamma_1 a_1 + \gamma_2 a_2 + (\gamma_3 - 1)a_3 \quad (4.22c)$$

Entsprechend ergeben sich für einen allgemeinen 3-Tor-Reihenadaptor die folgenden Gleichungen

$$b_1 = (1 - \gamma_1)a_1 - \gamma_1 a_2 - \gamma_1 a_3 \quad (4.23a)$$

$$b_2 = -\gamma_2 a_1 + (1 - \gamma_2)a_2 - \gamma_2 a_3 \quad (4.23b)$$

$$b_3 = -\gamma_3 a_1 - \gamma_3 a_2 + (1 - \gamma_3)a_3 \quad (4.23c)$$

Bild 4.3 zeigt die Signalfußdiagramme, die nach (4.22) und (4.23) direkt realisiert werden. Es ist klar zu sehen, daß für jeden Ausgang die Summe von drei Gleitkommazahlen durchzuführen ist. Daher betrachten wir nur einen Ausgang, z.B. b_1 , gemäß (4.22a). Die Berechnung aller anderen Ausgänge läuft auf ähnliche Weise.

Unter der Annahme, daß die Multipliziererkoeffizienten $(\gamma_1 - 1)$, γ_2 und γ_3 aus dem in (4.10a) gegebenen Gleitkommasystem GK, die Signale a_1 , a_2 und a_3 aus dem in (4.10b) gegebenen Gleitkommasystem sind und alle Multiplikationen ohne Quantisierungsfehler als Gleitkommazahlen aus dem in (4.11) gegebenen System dargestellt werden, betrachten wir die Berechnung des Ausgangssignals b_1 (Bild 4.3a) mit einem Akkumulator der Länge von $t_{\text{man}} + t'_{\text{man}}$ Bits (ohne Vorzeichenbit) plus zwei extra Bits als Schutzbits gegen Mantissenüberläufe.

Nun läßt sich b_1 auf folgende Weise

$$b_{10} = x_1 + x_2 + x_3$$

schreiben, wobei $x_1 = (\gamma_1 - 1)a_1 = m_1 2^{e_1}$, $x_2 = \gamma_2 a_2 = m_2 2^{e_2}$ und $x_3 = \gamma_3 a_3 = m_3 2^{e_3}$ ist. Ohne Verletzung der Allgemeinheit nehmen wir an, daß bei einer Summation von drei Gleitkommazahlen die folgenden Schritte auszuführen sind:

- 1) lösche den Akkumulator,
- 2) berechne x_1 exakt, d.h. mit $t_{\text{man}} + t'_{\text{man}}$ Bits für die Mantisse und akkumuliere es,
- 3) berechne x_2 exakt,
- 4) addiere $x_1 + x_2$ im Akkumulator auf,
- 5) berechne x_3 exakt,
- 6) addiere $x_3 + (x_1 + x_2)$ im Akkumulator auf,
- 7) quantisiere das Endergebnis, so daß die Bedingung (4.5) erhalten bleibt.

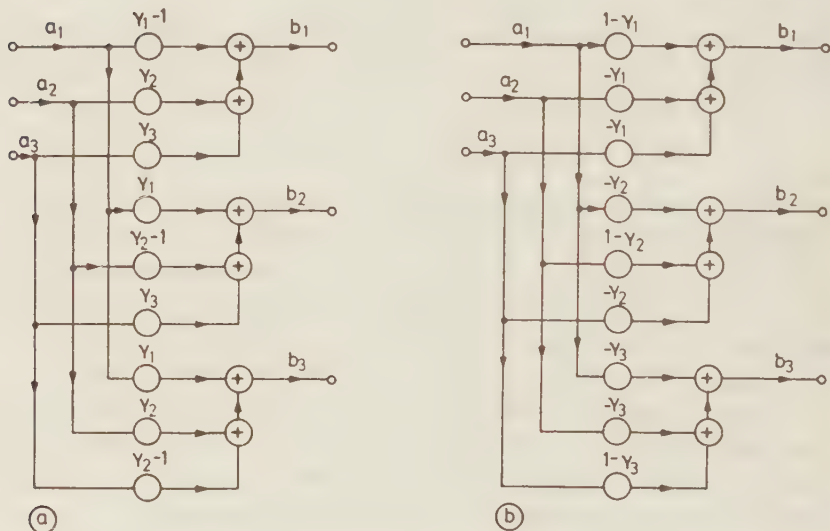


Bild 4.3: a) Signalflußdiagramm eines allgemeinen 3-Tor-Paralleladaptors nach (4.27),
 b) Signalflußdiagramm eines allgemeinen 3-Tor-Reihenadaptors nach (4.28).

Beim vierten Schritt haben wir gesehen, daß die Addition von zwei Gleitkommazahlen x_1 und x_2 im allgemeinen gemäß (4.16) gegeben werden kann. Zum vierten und sechsten Schritt nehmen wir an, daß $e_1 > e_2$ bzw. $e_1 > e_3$ gilt. Daher muß die Mantisse m_3 um $e_1 - e_3$ Binärstellen nach rechts geschoben werden. Diese Schiebeoperation macht aus x_3 eine neue Zahl x'_3 , die sich durch die folgende Beziehung schreiben läßt:

$$x'_3 = x_3 - \epsilon'_{abs} \quad (4.24)$$

Dabei ist ϵ'_{abs} durch (4.15) gegeben. Addiert man x'_3 zum Ergebnis des vierten Schritts, so erhält man für den mit $(t_{man} + t'_{man} + 2)$ Bits berechneten Adaptorausgang b_{11} die folgende Beziehung

$$b_{11} = x_1 + x_2 + x_3 - \epsilon_{abs} - \epsilon'_{abs} = m_{11} 2^{e_1} \quad (4.25)$$

Wir betrachten nun die zwei möglichen Fälle des Ergebnisses b_{11} .

a) $b_{10} > 0$

In diesem Fall ist einfach aus (4.25) zu ersehen, daß die folgende Beziehung gilt:

$$b_{11} \leq b_{10} \quad \text{und damit} \quad |b_{11}| \leq |b_{10}|$$

Schneidet man die Mantisse m_{11} auf die ursprüngliche Wortlänge t_{man} zurück, so erhält man den Adaptorausgang b_1 , der die Bedingung (4.5) erfüllt.

2) $b_{10} = x_1 + x_2 + x_3 < 0$.

Da ϵ_{abs} und ϵ'_{abs} stets positiv sind (vgl. (4.15)), können wir der Gleichung (4.25) entnehmen, daß für eine negative Summe die folgende Beziehung gilt:

$$|b_{11}| > |b_{10}|$$

Daraus folgt, daß das berechnete Signal b_1 betragsmäßig die Bedingung (4.5) nicht erfüllen kann. Um dies zu verhindern, addieren wir einen Korrekturterm der Größe $2^{(-t+1+e_1)}$ an den beiden Seiten von (4.25). Damit erhält man

$$b'_{11} = b_{11} + 2^{-t+1+e_1} = x_1 + x_2 + x_3 + (2^{-t+1+e_1} - \epsilon_{abs} - \epsilon'_{abs}) \quad (4.26)$$

Nach dem Schneiden des Betrags der resultierenden Mantisse $(m_{11} + 2^{-t+1})$ zurück auf t_{man} Binärstellen läßt sich die folgende Beziehung schreiben:

$$|b_1| = |[b_{11} + 2^{-t+1+e_1}]_q| < |x_1 + x_2 + x_3| = |b_{10}| \quad (4.27)$$

Aus den soeben deutlich diskutierten Fällen kann man ersehen, daß die Pseudopassivitätsbedingung (4.5) ebenso für einen allgemei-

nen 3-Tor-Adaptor eingehalten werden kann. Tatsächlich hat der Korrekturterm $(2^{-t+1}e_1)$ sehr geringe Auswirkung auf das Signal-Rausch-Verhältnis am Ausgang des Filters. Dieselbe Aussage gilt auch für den Korrekturterm (2^{-t+e_1}) beim 2-Tor-Adaptor.

Für einen 3-Tor-Paralleladaptor, dessen Tor 3 reflexionsfrei

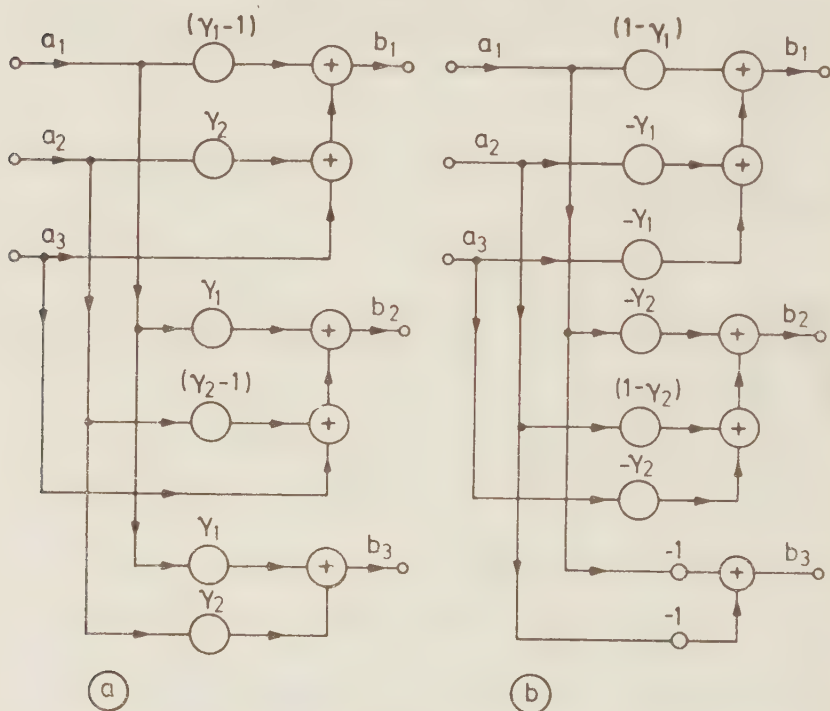


Bild 4.4: Signalflußdiagramme für einen 3-Tor-Adaptor mit $\gamma_3=1$
a) für einen Paralleladaptor, b) für einen Reihenadaptor.

ist, lassen sich aus (2.16) die folgenden Gleichungen aufstellen:

$$b_1 = (\gamma_1 - 1)a_1 + \gamma_2 a_2 + a_2 \quad , \quad (4.28a)$$

$$b_2 = \gamma_1 a_1 + (\gamma_2 - 1)a_2 + a_3 \quad , \quad (4.28b)$$

$$b_3 = \gamma_1 a_1 + \gamma_2 a_2 \quad . \quad (4.28c)$$

Das zugehörige Signalflußdiagramm (Bild 4.4a) erfordert sechs Multiplikationen und fünf Additionen. Entsprechend ergeben sich für die 3-Tor-SerienadapTORausgänge die Gleichungen

$$b_1 = (1-\gamma_1)a_1 - \gamma_1 a_2 - \gamma_1 a_3, \quad (4.29a)$$

$$b_2 = -\gamma_2 a_1 + (1-\gamma_2)a_2 - \gamma_2 a_3, \quad (4.29b)$$

$$b_3 = -a_1 - a_2. \quad (4.29c)$$

Das Signalflußdiagramm in Bild 4.4b zeigt, daß zur Realisierung ebenfalls sechs Multiplikationen und fünf Additionen erforderlich sind.

Wie man aus (4.28) und (4.29) ersehen kann, erfolgen die Berechnungen von b_1 und b_2 genau wie bei einem allgemeinen 3-Tor-Adaptor und die Berechnung von b_3 wie bei einem 2-Tor-Adaptor.

Wir können die vorstehenden Ergebnisse so zusammenfassen, daß die Pseudopassivitätseigenschaften eines 2-Tor-Adaptors mit der Addition von einem LSB (2^{-t}) zur berechneten negativen Mantisse garantiert wird. Für einen allgemeinen 3-Tor-Adaptor gilt dieselbe Aussage mit der Ausnahme, daß eine Addition von zwei LSB ($2 \cdot 2^{-t} = 2^{-t+1}$) zur negativen Mantisse erforderlich ist.

4.3.3 Pseudopassivitätseigenschaften des Übertragers

Im folgenden wollen wir die Pseudopassivitätseigenschaften der in Tabelle 2.3 aufgeführten Realisierungen eines idealen Übertragers untersuchen. Für die Realisierungsmöglichkeit a (Tabelle 2.3) bleibt die Pseudoenergieneutralität nur dann erhalten, wenn die Multipliziererkoeffizienten n und $1/n$ exakt reziprok zueinander realisiert werden [26,62]. Dies ist aber nur möglich, wenn n und $1/n$ als Gleitkommazahlen aus dem Gleitkommasystem

$$GK = GK(2, 1, e_{\min 1}, e_{\max 1}) \quad (4.30)$$

dargestellt werden (siehe Abschnitt 3.7).

Unter der Voraussetzung, daß die Signale a_1, b_1, a_2 und b_2 als Gleitkommazahlen aus (4.10b) in Zweierkomplementkode dargestellt sind und daß n eine Zweierpotenz ist, lassen sich die nicht normierten Mantissen der Multiplikationen $(1/n)a_2$ und $n a_1$ mit $t_{\text{man}} + 1$ Bits (ohne Vorzeichenbit) exakt darstellen. Es sei bemerkt, daß ein Renormierungsschritt die Wortlänge der resultierenden Mantisse auf t_{man} Bits zurückbringt. Nach diesem Renormierungsschritt kann man leicht nachprüfen, daß die Pseudopassivitätsbedingung (4.5) stets erfüllt wird, solange kein Exponentenunterlauf auftritt. Dieselbe

Aussage gilt auch in dem Fall, wenn der Exponentenunterlauf als Null weiterbehandelt wird. Das bedeutet, daß der Übertrager (Tabelle 2.4b) durch Verwendung der Quantisierungskennlinien (vgl. Bild 4.9b) nach den Multiplikationen einfach pseudopassiv bleibt. Wird beim Exponentenunterlauf automatisch auf die Renormalisation verzichtet, so muß ein Korrekturterm der Größe $2^{-t_{\text{man}}}$ zur negativen Mantissee addiert werden, um die Bedingung (4.5) einzuhalten. Um dies näher zu erläutern, nehmen wir an, daß z.B. das Signal $a_2 = m_2 \cdot 2^2$ und $n = 0.1 \cdot 2^e$ ist, wobei m_2 eine Wortlänge von t_{man} besitzt. Die exakte Multiplikation na_2 läßt sich als

$$b_{10} = m'_2 2^{e+e_2} \quad (4.31)$$

ausdrücken. Damit ist m'_2 eine nicht normierte Mantissee der Länge von $t_{\text{man}} + 1$ Bits. Sei $e+e_2 < e_{\text{min}}$, so wird in diesem Fall m'_2 um $|e+e_2 - e_{\text{min}}|$ Binärstellen nach rechts geschoben. Damit erhalten wir das mit t_{man} Bits für die Mantissee berechnete Signal

$$b'_1 = b_{10} \cdot 2^{-\epsilon_{\text{abs}}} = m'_1 2^{e_{\text{min}}} \quad (4.32)$$

wobei ϵ_{abs} durch

$$0 \leq \epsilon_{\text{abs}} < 2^{-t_{\text{man}} + e_{\text{min}}} \quad (4.33)$$

gegeben ist.

Falls b_{10} positiv ist, kann man leicht aus (4.32) ersehen, daß die Bedingung

$$|b_1| = |b'_1| \leq |b_{10}|$$

erfüllt ist. Wenn b_{10} negativ ist, erhalten wir unter Verwendung von (4.32) und (4.33)

$$b_1 = b'_{1+2} \cdot 2^{-t_{\text{man}} + e_{\text{min}}} = (m'_{1+2} \cdot 2^{-t_{\text{man}}}) 2^{e_{\text{min}}} = b_{10} \cdot (2^{-t_{\text{man}} + e_{\text{min}} - \epsilon_{\text{abs}}})$$

Daraus folgt

$$|b_1| < |b_{10}|$$

Dieselbe Aussage gilt auch für das Signal b_2 .

Für die Realisierung gemäß Tabelle 2.4c kann man leicht nach-

prüfen, daß die Pseudopassivität stets erfüllt ist, solange die vorher diskutierten Maßnahmen für die Erfüllung der Pseudopassivität eines 2-Tor-Adaptors in Betracht gezogen werden. Im Fall der Realisierung gemäß Tabelle 2.4d sind die Koeffizienten n und $1/n$ aus Realisierungsgründen Potenzen von 2 in der Form 2^e , wobei e eine ganze Zahl ist. Dann wird der ideale Übertrager beschrieben durch

$$b_1 = -\gamma a_1 + \left(\frac{1+\gamma}{n}\right) a_2 = -\gamma a_1 + \gamma'_2 a_2 \quad (4.34a)$$

$$b_2 = \gamma_1 n a_1 + \gamma a_2 = \gamma'_1 a_1 + \gamma a_2 \quad (4.34b)$$

Bild 4.5 zeigt das zugehörige Wellenflußdiagramm. Diesem Bild kann man entnehmen, daß diese Realisierung und diejenige, die in Bild 4.2 für einen 2-Tor-Adaptor gezeigt ist, mit γ'_1 und γ'_2 anstatt γ_1 und γ_2 ähnlich sind. Es sei bemerkt, daß die Mantissen von γ'_1 und γ_1 dieselbe Wortlänge t'_{man} besitzen.

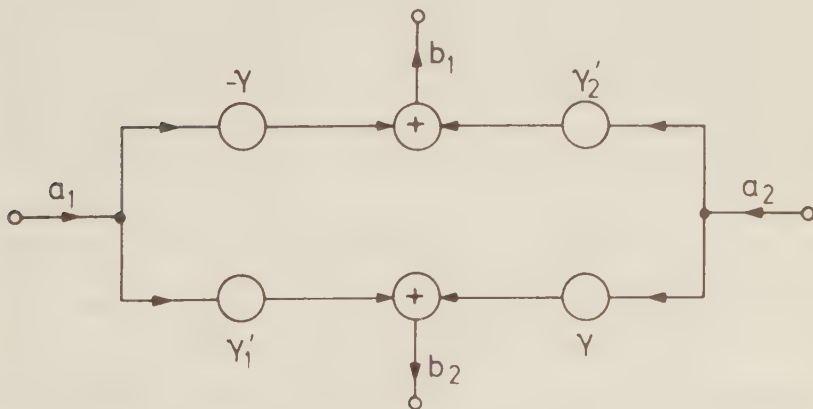


Bild 4.5: Wellenflußdiagrammrealisierung nach (4.34).

Der Grund besteht darin, daß die Multiplikation von γ_1 mit n nur Erhöhung oder Erniedrigung des Exponenten bedeutet, da, wie schon erwähnt wurde, n und $1/n$ Potenzen von 2 in der Form 2^e sind. Für γ'_2 und γ_2 gilt auch dieselbe Aussage.

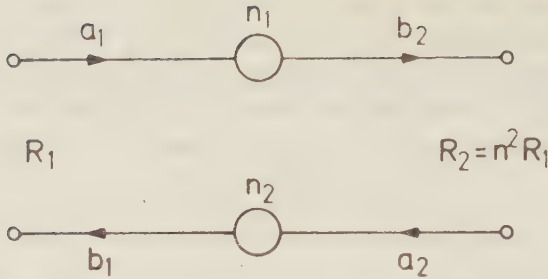


Bild 4.6: Wellenflußdiagramm eines realen Übertragers.

Unter der Annahme, daß die Multiplikationen γa_1 , γa_2 , $\gamma_1^* a_1$ und $\gamma_2^* a_2$ exakt durchgeführt werden, kann man einfach feststellen, daß die Pseudopassivität eines idealen Übertragers gemäß Bild 4.5 durch die diskutierten Maßnahmen für einen 2-Tor-Adaptor stets sichergestellt wird.

In anderen Realisierungsfällen werden im Interesse der Stabilität und auch der Realisierungsmöglichkeit die Koeffizienten so quantisiert, daß der Übertrager zumindest pseudopassiv bleibt. Man spricht in diesem Fall vom "realen" Übertrager [29]. Bild 4.6 zeigt das zugehörige Wellenflußdiagramm. Um eine Bedingung für die Koeffizienten herleiten zu können, gehen wir von der Pseudoleistung p aus, die von diesem Übertrager aufgenommen wird. Dann erhalten wir [29]

$$\begin{aligned}
 p &= (a_1^2 - n_2^2 a_2^2) G_1 + (a_2^2 - n_1^2 a_1^2) G_1 / n^2 \\
 &= (1 - (\frac{n_1}{n})^2) G_1 a_1^2 + (\frac{1}{n^2} - n_2^2) G_1 a_2^2 \quad . \quad (4.35)
 \end{aligned}$$

Aus Pseudopassivitätsgründen soll p nicht negativ werden. Daraus folgt, daß die Quantisierung der Koeffizienten derart erfolgt, daß die Bedingung

$$n_1 \leq n \leq \frac{1}{n_2} \quad (4.36)$$

erfüllt ist.

Setzen wir voraus, daß die Signale und die Koeffizienten n_1 und n_2 unterschiedliche Wortlänge besitzen, so erhalten wir

$$b_1 = b_{10}(1-\epsilon_1) \quad , \quad b_2 = b_{20}(1-\epsilon_2) \quad , \quad (4.37)$$

wenn die resultierende Mantisse nach der Multiplikation auf t_{man} Binärstellen zurückgeschnitten wird. Verwendet man Betragsschneiden als Quantisierungsmaßnahme, so sind die in (4.37) gegebenen ϵ_1 und ϵ_2 stets positiv. Daraus folgt

$$|b_u| \leq |b_{u0}| \quad , \quad u = 1, 2 \quad ,$$

und somit ist die Pseudopassivität des Übertragers eingehalten. Dies gilt immer, solange der Exponentenunterlauf als Null weiterbehandelt wird. Verzichtet man beim Exponentenunterlauf automatisch auf die Normalisation, so muß, wie schon vorher erwähnt wurde, ein Korrekturterm der Größe $2^{-t_{\text{man}}}$ zur negativen Mantisse addiert werden, um die Pseudopassivitätsbeding (4.5) zu erfüllen.

4.4 Inkrementale Pseudopassivität der Adaptoren und der Übertrager

Wie schon erwähnt wurde, muß die Signalwortlänge der Adaptor- und Übertragerausgänge wieder auf die ursprüngliche Signalwortlänge passend reduziert werden, um die Berechnung mit wachsender Wortlänge zu vermeiden. Diese Reduzierung der Wortlänge wird mit Hilfe geeigneter nichtlinearer Operationen ausgeführt. Durch die vorherigen Diskussionen können wir davon ausgehen, daß diese Nichtlinearitäten als nichtlineare Eintore, die mit Hilfe von 3-Tor-Zirkulatoren in die Ausgänge der Adaptoren und Übertrager eingefügt sind [29,30], betrachtet werden. Bild 4.7 zeigt ein allgemeines n-Tor-Bauelement eines WDF's, bei dem die Wirkungen der nichtlinearen Operationen in den nichtlinearen Eintoren $Q_u(\cdot)$ ($u=1,2,\dots,n$) zusammengefaßt sind. Es sei bemerkt, daß das in Bild 4.7 angegebene Wellendigitalbauelement N entweder ein einzelnes Element (wie z.B. Adaptor, Übertrager) eine Zusammenschaltung aus mehreren Elementen oder auch ein vollständiges Wellendigitalfilter darstellt.

Zur Beurteilung der inkrementalen Pseudopassivität des n-Tors N betrachten wir nur die nichtlinearen Eintore, da N_0 inkremental pseudopassiv ist [29]. Um den Einfluß der nichtlinearen Operationen zu untersuchen, liegt es nahe, das inkrementale Pseudopassivitätsverhalten eines einzelnen nichtlinearen Eintores zu diskutieren. Wir betrachten das Eintor, welches durch $Q_u(\cdot)$ beschrieben ist (Bild 4.8).

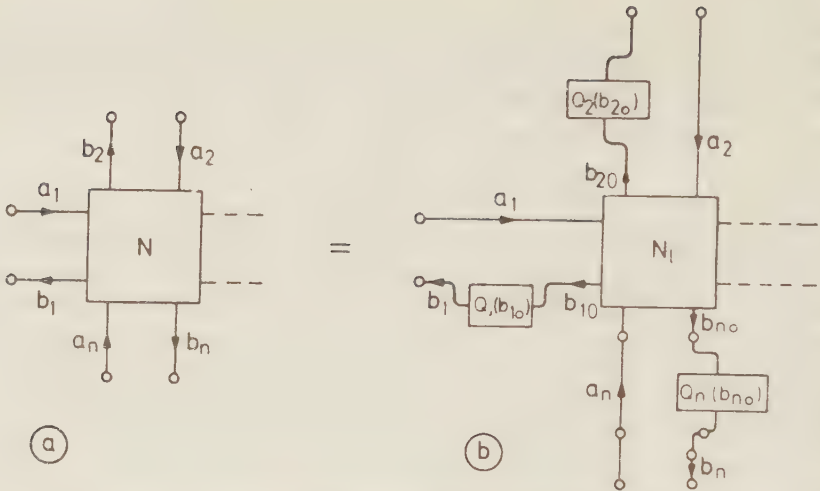


Bild 4.7: a) Ein nichtlineares n -Tor-Wellendigitalbauelement, b) eine zugehörige Anordnung, die aus dem entsprechenden linearen Element N_l , n 3-Tor-Zirkulatoren und n nichtlinearen Elementen mit den Quantisierungsfunktionen $Q_v(\cdot)$ ($v=1, 2, \dots, n$) besteht.

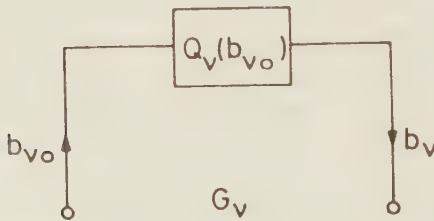


Bild 4.8: Ein nichtlineares Eintor mit der Quantisierungsfunktion $Q_v(b_{v0})$.

Wie man aus Bild 4.8 ersehen kann, wird der Zusammenhang zwischen dem idealen linearen Ausgangssignal b_{v0} und dem nichtlinearen b_v durch die Quantisierungsfunktion $Q_v(\cdot)$ ($v=1, 2, \dots, n$) beschrieben. Dann erhalten wir

$$b_v = Q_v(b_{v0}) \quad . \quad (4.38)$$

Die Pseudoleistung, die von dem Eintor (Bild 4.8) aufgenommen

wird, ist gegeben durch

$$P = (b_{U0}^2 - Q_U^2(b_{U0}))G_U \quad .$$

Wie schon vorher diskutiert wurde, ist das Eintor pseudopassiv, falls die Bedingung

$$|Q_U(b_{U0})| = |b_U| \leq |b_{U0}| \quad (4.39)$$

erfüllt wird. Es ist klar zu sehen, daß (4.39) offenbar durch die vorher gegebenen Maßnahmen sichergestellt werden kann.

Nun betrachten wir zur Beurteilung der inkrementalen Pseudopassivität des Eintors zwei verschiedene Eingangssignale b_{U0} und b'_{U0} . Die zugehörigen Ausgangssignale b_U und b'_U sind durch

$$b_U = Q_U(b_{U0}) \quad , \quad b'_U = Q_U(b'_{U0})$$

gegeben. Die inkrementale Pseudoleistung P_Δ läßt sich durch

$$P_\Delta = [(b_{U0} - b'_{U0})^2 - (Q_U(b_{U0}) - Q_U(b'_{U0}))^2]G_U \quad (4.40)$$

beschreiben. Da das Eintor keine Verzögerungen enthält [29], wird die inkrementale Pseudopassivität des Eintores eingehalten, falls die Bedingung $P_\Delta \geq 0$ für alle $t_m > 0$ und für alle möglichen Eingangssignale erfüllt ist. Daraus folgt

$$|b_{U0} - b'_{U0}| \geq |Q_U(b_{U0}) - Q_U(b'_{U0})| \quad . \quad (4.41)$$

Berücksichtigt man die Granularität der Signale b_U und b_{U0} , so kann man sich davon leicht überzeugen, daß eine nichtkonstante Quantisierungsfunktion $Q_U(\cdot)$ die Bedingung (4.41) nicht erfüllen kann. Wird aber der Einfluß der Granularität der Signale vernachlässigt, so läßt sich die inkrementale Pseudopassivität durch die in Bild 4.9 angegebenen Quantisierungsfunktionen $Q_U(\cdot)$ erreichen. Diese Vernachlässigung der Granularität rechtfertigt die Behauptung, daß durch die Granularität hervorgerufenen Fehler wesentlich kleiner sind als diejenigen, die durch einen Exponentenüberlauf entstehen können. Dies gilt, wenn die Mantissenwortlänge hinreichend groß ist. Bild 4.10 zeigt die Feinstruktur der Quantisierungsfunktion $Q_U(b_{U0})$ gemäß Bild 4.9a.

Als Zusammenfassung der in vorherigen Unterabschnitten diskutierten Maßnahmen sei bemerkt, daß bei Verwendung einer Zweierkomplement-Gleitkomma-Arithmetik zur Erhaltung der sowohl ge

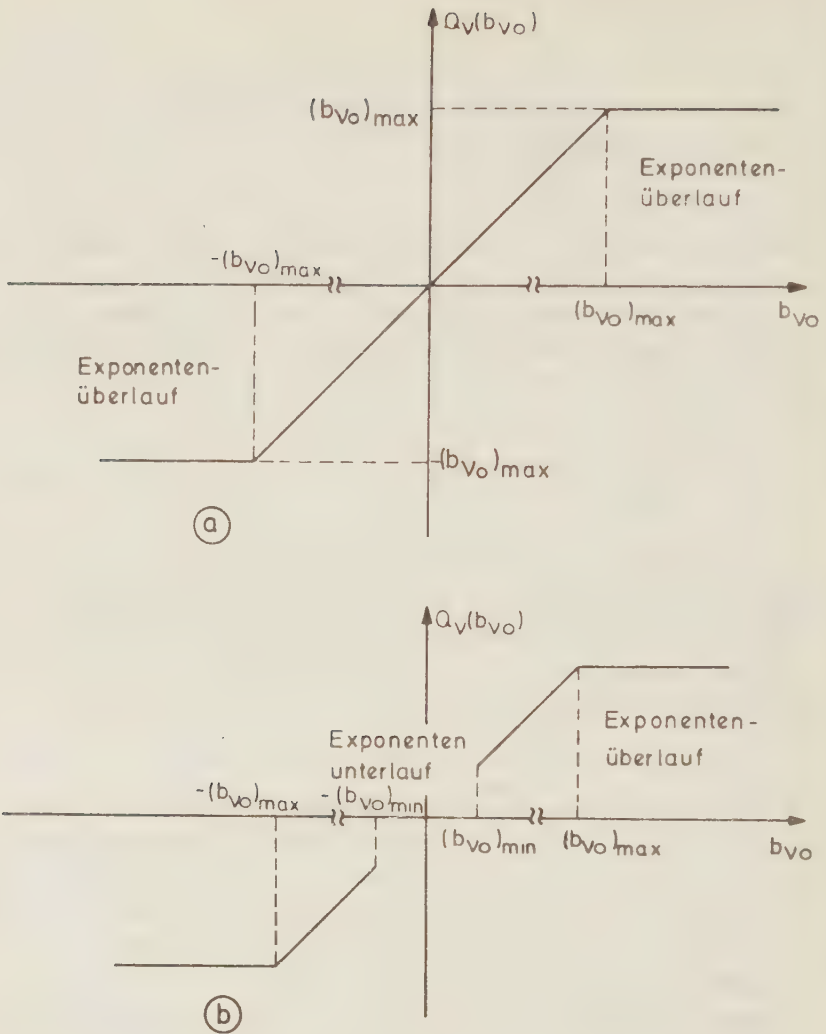


Bild 4.9: Quantisierungskennlinien der Funktion $Q_v(b_{vo})$ ($v=1,2,\dots$) mit Sättigungscharakteristik beim Exponentenüberlauf und beim Exponentenunterlauf:
a) wird automatisch auf die Normierung verzichtet,
b) wird durch Null ersetzt (ohne Feinstruktur).

wöhnlichen als auch der inkrementalen Pseudopassivität eines elementaren (allgemeinen) Adaptors die Erhöhung der Signalwortlänge inner

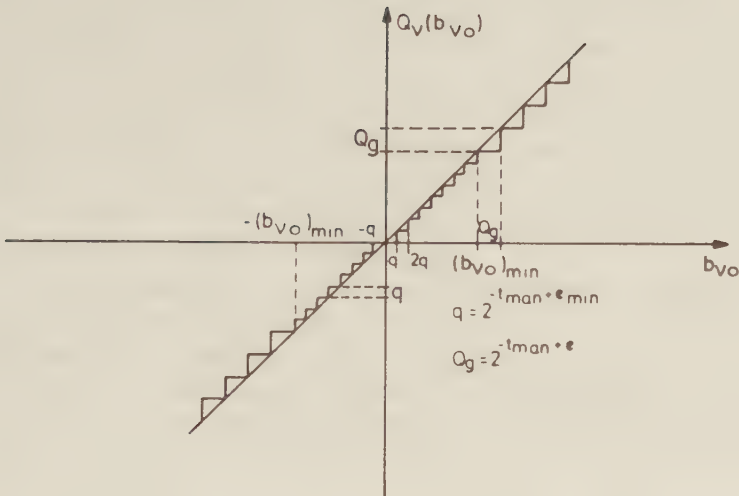


Bild 4.10: Feinstruktur der Quantisierungsfunktion $Q_V(b_{VO})$,
(gemäß Bild 4.9a) hervorgerufen durch Betragsschneiden.

halb des Adaptors vorgenommen werden muß. An den Adaptorausgängen benötigen wir zusätzliche Korrekturterme, die zu den negativen Ausgängen hinzuzuaddieren sind. Zur Reduzierung der Signalwortlänge dient in diesem Fall das Betragsschneiden. Selbstverständlich wird durch die Erhöhung der Wortlänge ein zusätzlicher Realisierungsaufwand notwendig.

Im folgenden wollen wir die Stabilitätseigenschaften einer speziellen Wellendigitalfilterstruktur, die nur mit 2-Tor-Adaptoren realisiert wird, untersuchen.

4.5 Stabilitätseigenschaften von Bireziprok-Wellendigitalfiltern

Eine besondere Struktur von Brücken-Wellendigitalfiltern ist gegeben, falls die Dämpfung der Reflektanz S_{11} spiegelbildlich zu der Transmittanz S_{21} ist, bezogen auf ein Viertel der tatsächlichen Betriebsrate $F=1/T$. Man nennt die Struktur in diesem Fall Bireziprok-Wellendigitalfilter. Diese Struktur hat einen sehr stark reduzierten Realisierungsaufwand [56].

Die Realisierung der Reflektanz S_{11} und/oder der Transmittanz S_{21} ist auf das Problem reduziert worden, die Reflektanz S_1 und S_2 bzw. die kanonischen Impedanzen beider Brückenarme zu realisieren.

ren. Dafür gibt es zahlreiche Möglichkeiten [19,20,54,55]. Wir betrachten die Realisierung der beiden Zweige eines Brücken-Wellendigitalfilters als Kaskadenschaltung von Allpässen erster und zweiter Ordnung, die über 3-Tor-Zirkulatoren zusammengeschaltet sind [2]. Da die hierbei auftretenden 3-Tor-Zirkulatoren den gleichen Zirkulationswiderstand aufweisen, benötigt man für die Zusammenschaltung keine zusätzlichen Adaptoren. Bild 4.11 zeigt das Wellenflußdiagramm eines Bireziprok-Wellendigitalfilters [58].

Aus Bild 4.11 ist ersichtlich, daß bei abgeschaltetem Eingangssignal ($a_1=0$) die Untersuchung des in Bild 4.12 dargestellten Allpassgliedes ausreicht, um das Verhalten des gesamten Filters zu untersuchen. Nun wollen wir uns der Frage zuwenden, inwieweit die bereits erwähnte Pseudopassivität eines Allpassgliedes bei abgeschaltetem Eingangssignal und beliebigen Anfangszuständen eingehalten werden kann.

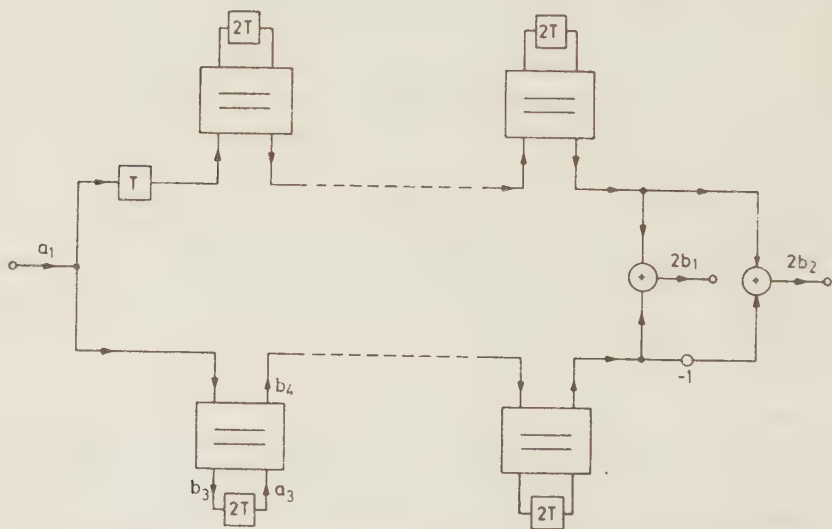


Bild 4.11: Wellenflußdiagramm eines bireziproken Wellendigitalfilters.

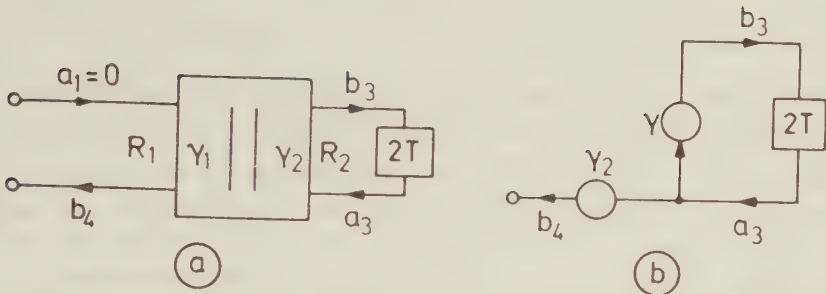


Bild 4.12: a) Ein Allpassglied ersten Grades eines Bireziprok-Wellendigitalfilters mit $a_1=0$,

b) die zugehörige Realisierung $(\gamma = \frac{R_1 - R_2}{R_1 + R_2})$.

Anhand der in Bild 4.12b dargestellten Schaltung erhalten wir

$$b_3 = \gamma a_3 \quad (4.42)$$

Nimmt man an, daß der Exponentenunterlauf durch Null behandelt wird, so gilt im Fall des Betragsschneidens nach der Multiplikation die folgende Bedingung

$$|b_3| \leq |b_{30}| = |\gamma a_3| \quad (4.43)$$

Wegen $|\gamma| < 1$ gilt stets

$$|b_3| < |a_3|$$

und daraus folgt, daß die beiden Signale a_3 und b_3 gegen Null konvergieren. Somit gilt dieselbe Aussage sukzessiv für alle weiteren Allpaßglieder eines Brückenzeiges, so daß schließlich auch das Ausgangssignal des gesamten Filters gegen Null strebt.

Wird beim Exponentenunterlauf auf die Normierung verzichtet, so muß zu der nach der Multiplikation resultierenden negativen Mantisse ein LSB der Größe $2^{-t_{\text{man}}}$ addiert werden, wie schon in vorherigen Unterabschnitten erwähnt wurde, um die Bedingung (4.43) zu erfüllen.

Im Fall des Zweierkomplementschneidens ist es einfach nachzuprüfen, daß sich die Bedingung (4.43) durch das Endschneiden wie folgt darstellen läßt:

$$|b_3| < |\gamma a_3|, \text{ falls } b_3 > 0, \quad (4.44a)$$

$$|b_3| = |[\gamma a_3 + 2^{-t_{\text{man}} + e}]_q| < |\gamma a_3|, \text{ falls } b_3 < 0, \quad (4.44b)$$

wobei e der Exponent von a_3 ist. Aus (4.44) ist ersichtlich, daß die beiden Signale a_3 und b_3 im Fall des Endscheidens durch eine Addition von 1 LSB ($2^{-t_{\text{man}}}$) zur negativen Mantisse gegen Null konvergieren. In diese Richtung strebt das Ausgangssignal des gesamten Filters. Fassen wir die vorstehenden Ergebnisse zusammen, so können wir sagen, daß sich solche Filter, die aus Kaskadenschaltung von Allpaßgliedern erster Ordnung in z oder z^2 bestehen, bei abgeschaltetem Eingangssignal und beliebigen Anfangszuständen durch einfache Maßnahmen wie passive Schaltungen verhalten. Es sei an dieser Stelle ausdrücklich darauf hingewiesen, daß die Polyphasen-Digitalfilterbänke, die unter Verwendung von Wellendigitalfiltern realisiert werden [60], durch dieselben für Bireziprok-Wellendigitalfilter diskutierten Maßnahmen bei abgeschaltetem Eingangssignal pseudopassiv bleiben.

4.6 Filterbeispiele

Abschließend sollen noch einige Beispiele zur Demonstration der Stabilitätseigenschaften von Wellendigitalfiltern bei unterschiedlichen Betriebsbedingungen, im Hinblick auf das Auftreten parasitärer Schwingungen, vorgestellt werden. Die Untersuchungen wurden mit Gleitkommaarithmetik auf dem Rechner HP1000 durchgeführt. Es wird gezeigt, daß sich die vorigen theoretisch diskutierten Pseudopassivitätseigenschaften praktisch bestätigen lassen.

Zunächst betrachten wir das folgende Filter zur Untersuchung der Stabilität bei abgeschaltetem Eingangssignal. Es handelt sich hier um ein Brücken-Wellendigitalfilter 9. Grades [61], dessen Koeffizientenwerte in Gleitkommaarstellung wie folgt angegeben sind:

$$\gamma_1 = 0.10111 \cdot 2^{-1}, \quad \gamma_2 = 0.111 \cdot 2^0,$$

$$\gamma_3 = 0.111 \cdot 2^{-3}, \quad \gamma_4 = 0.10100001 \cdot 2^0.$$

Das Wellenflußdiagramm und der sich mit 14 Bits Mantissenwortlänge und 4 Bits für den Exponenten ergebende Dämpfungsverlauf sind in

Bild 4.13 dargestellt.

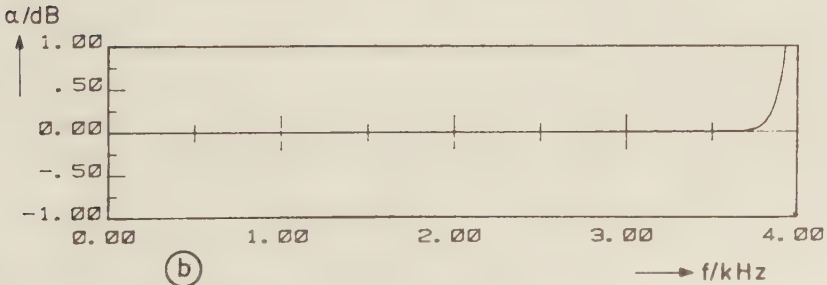
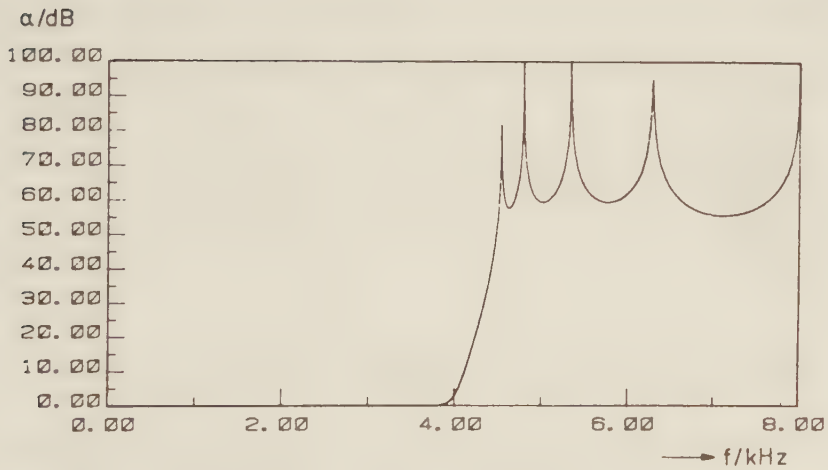
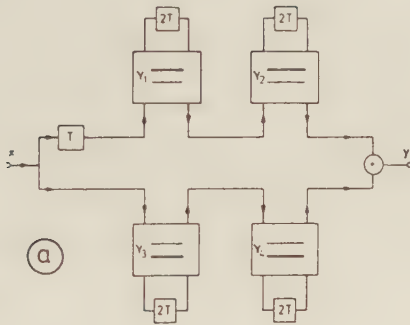


Bild 4.13: a) Wellenflußdiagramm des Brücken-Wellendigitalfilters, b) zugehöriger Dämpfungsverlauf.

Um möglichst ungünstige autonome Betriebsbedingungen zu simulieren (also mit einer Mantisse, die nur 5 Bit umfaßt), haben wir eine Signalwortlänge aus dem Gleitkommasystem GS(2,5,-4,3) gewählt. Es wurden dann, als Anfangszustand, alle Speicherwerte auf den negativen Maximalwert -8 gesetzt. Darüberhinaus ist Betragsschneiden nur an denjenigen Adaptorausgängen angewandt worden, die zu einem Verzögerer führen. Bild 4.14 zeigt den dafür erhaltenen zeitlichen Verlauf des Filterausgangs $y(mT)$. Wie zu erkennen ist, nimmt die Signalamplitude, mit ihren anfangs großen Werten, sehr schnell ab. Zur Interpretation des Ergebnisses sei bemerkt, daß trotz der groben Granularität die Stabilität bei abgeschaltetem Eingangssignal erzielt wurde.

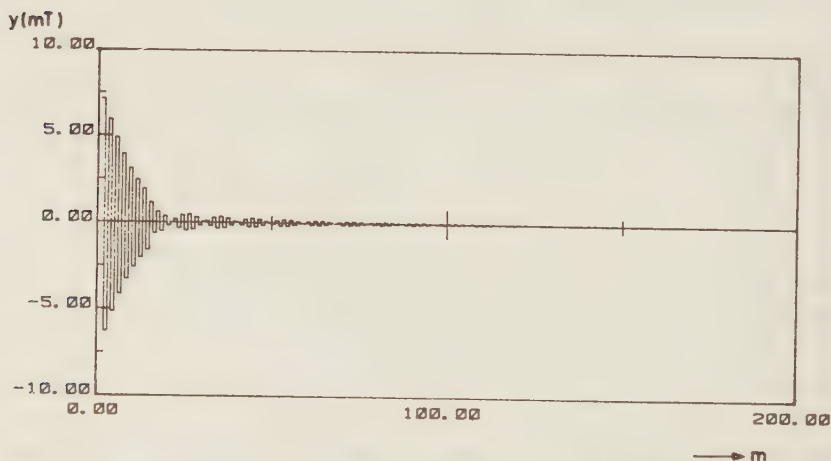


Bild 4.14: Ausgangssignal $y(mT)$ des in Bild 4.13 dargestellten Filters bei abgeschaltetem Eingangssignal.

Zur Untersuchung der Antwortstabilität betrachten wir zwei Filter, bei denen es sich um einen Bandpaß 14. Grades und um einen Tiefpaß 3. Grades handelt. Wir wenden uns zunächst dem Bandpaß 14. Grades [20] zu, dessen Koeffizientenwerte in Gleitkomma wie folgt dargestellt sind:

$$\begin{aligned} \gamma_1 &= -0.111 \cdot 2^{-1} & , & \quad \gamma_2 = 0.111 \cdot 2^{-2} & , & \quad \gamma_3 = 0.111 \cdot 2^{-2} \\ \gamma_4 &= 0.1 \cdot 2^0 & , & \quad \gamma_5 = 0.1101 \cdot 2^{-2} & , & \quad \gamma_6 = 0.111 \cdot 2^{-3} \\ \gamma_7 &= 0.1101 \cdot 2^{-1} & . & \end{aligned}$$

Das Wellenflußdiagramm und der mit dem Gleitkommasystem GS(2,14,-16,15) berechnete zugehörige Dämpfungsverlauf sind in Bild 4.15 gezeichnet.

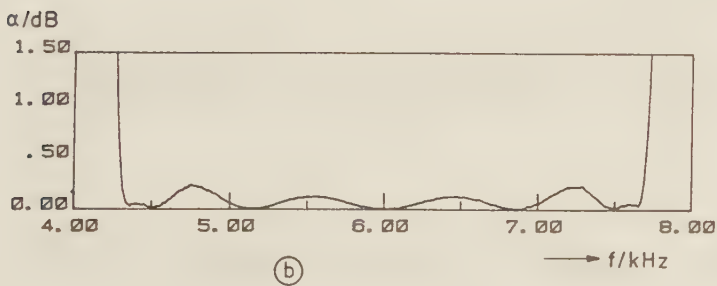
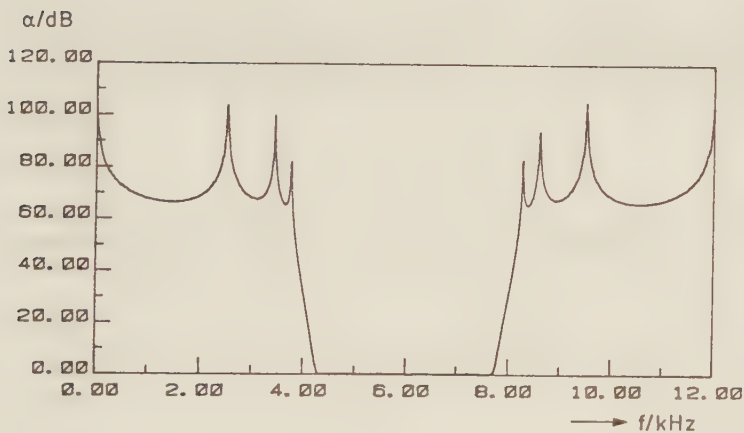
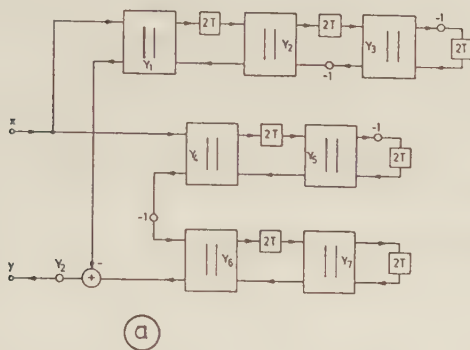


Bild 4.15: a) Wellenflußdiagramm des Filters 14. Grades,
b) zugehöriger Dämpfungsverlauf.

Für die folgende Untersuchung wurden die Signalparameter aus dem Gleitkommasystem GS(2,14,-16,1) gewählt. Verwendet wurde ein Eingangssignal $x(mT)$, welches einen exponentenüberlauffreien stationären Betrieb zuließ. Bild 4.16 zeigt die Signalverläufe des Signals $x(mT)$ und des zugehörigen Ausgangssignals $y_o(mT)$ für einen Ausschnitt mit 100 Abtastwerten nach einer Einschwingdauer von 50 Abtastwerten. Parallel hierzu wurde dasselbe Filter mit gleichem Eingangssignal simuliert, das aber nach Ablauf der Einschwingdauer gestört wurde. Diese Störung wurde dadurch verursacht, daß 19 aufeinanderfolgende Eingangswerte auf den Wert -2 gesetzt wurden

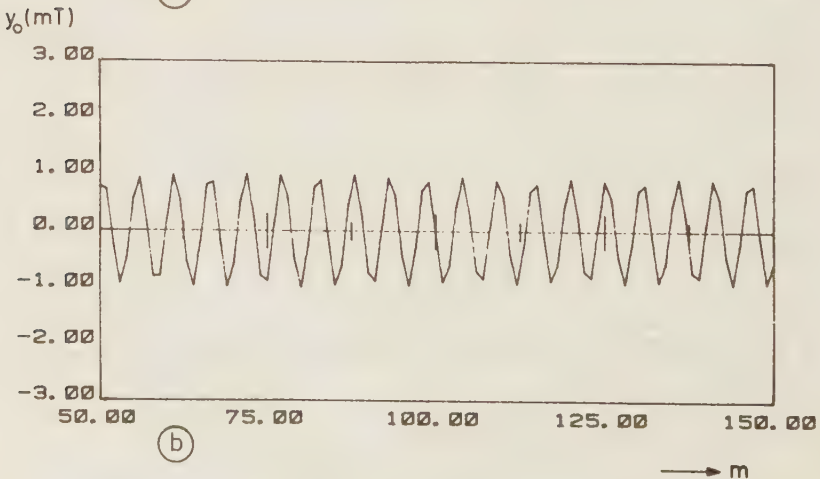
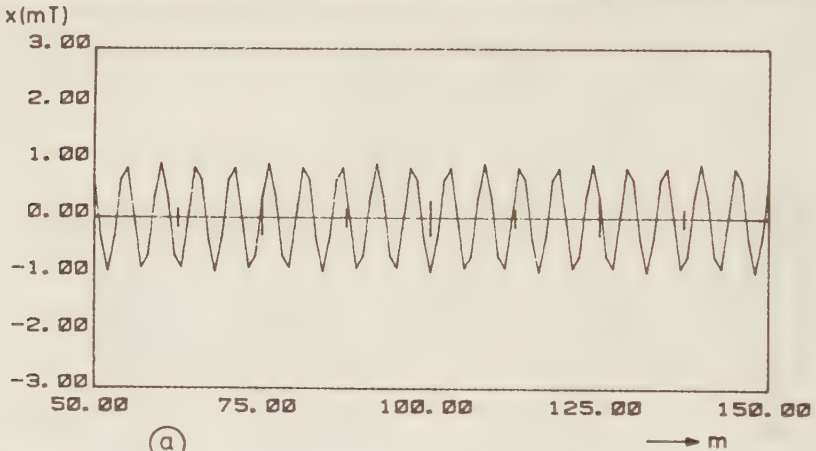


Bild 4.16: a) Ungestörtes Eingangssignal $x(mT)$,
b) zugehöriges Ausgangssignal $y_o(mT)$.

und es wurde überprüft, daß hierdurch tatsächlich Überläufe im Innern des Filters stattfanden. Bild 4.17 stellt die dafür erhaltenen zeitlichen Signalverläufe des Eingangs $x(mT)$ und des Ausgangs $y(mT)$ dar.

Nach Ende der Störung wurde das Differenzsignal

$$e(mT) = y_0(mT) - y(mT)$$

gebildet, dessen zeitlicher Verlauf in Bild 4.18 dargestellt ist. Wie man aus Bild 4.18 ersieht, konvergiert dieses Differenzsignal sehr rasch gegen Null. Tatsächlich wurde

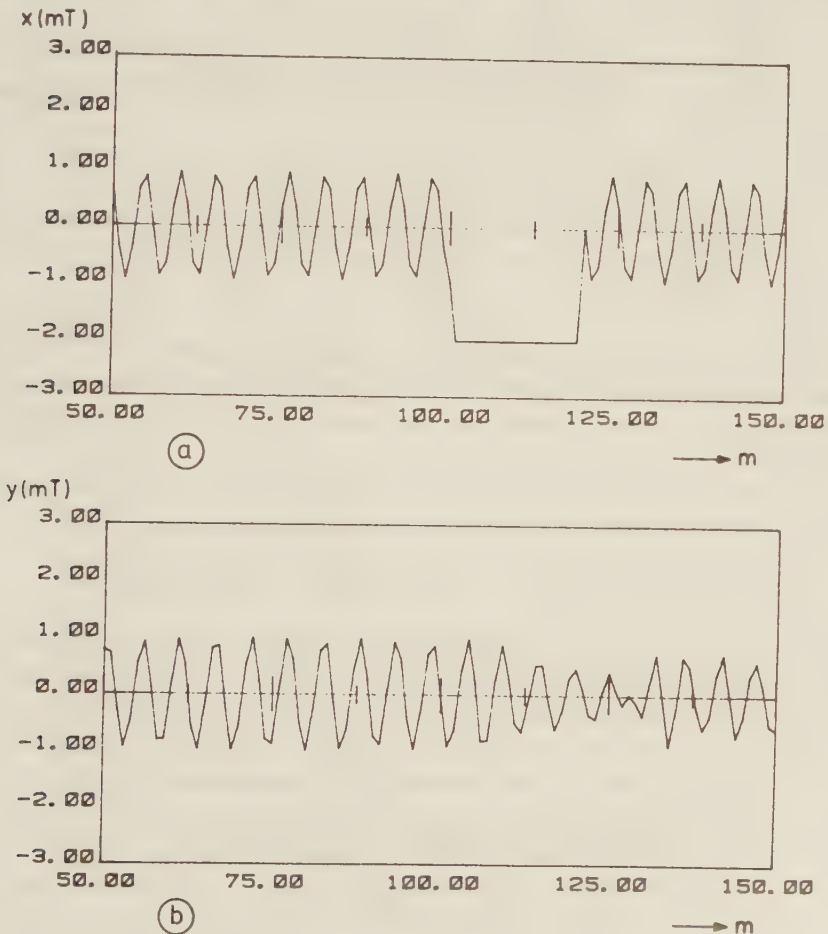


Bild 4.17: a) Gestörtes Eingangssignal,
b) zugehöriges Ausgangssignal $y(mT)$.

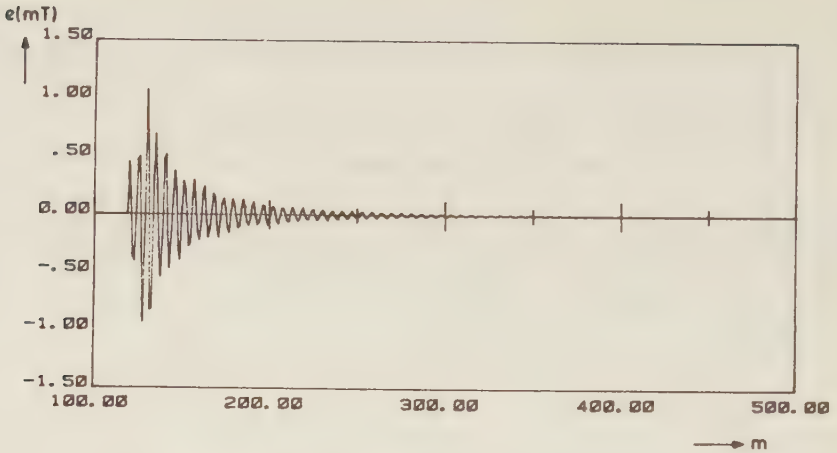


Bild 4.18: Differenzsignal $e(mT)$.

$$e(mT) = 0 \quad \text{für} \quad m > 450$$

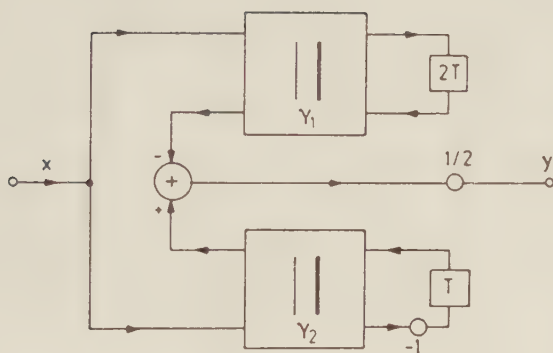
erreicht. Das bedeutet, daß das Ausgangssignal $y(mT)$ mit gestörtem Eingangssignal, trotz intern aufgetretener Exponentenüberläufe, in endlicher Zeit wieder gegen das Ausgangssignal des ungestörten Filters konvergiert.

Als nächstes betrachten wir den Tiefpaß 3. Grades (vgl. Bild 4.19), dessen Koeffizientenwerte wie folgt angegeben sind:

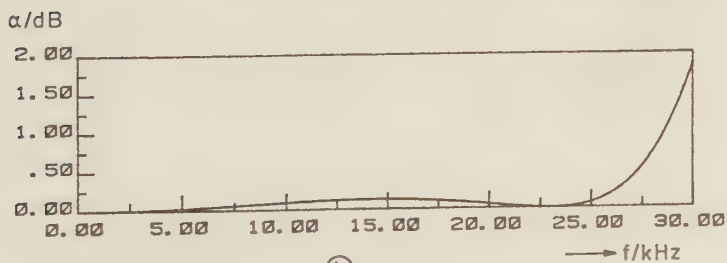
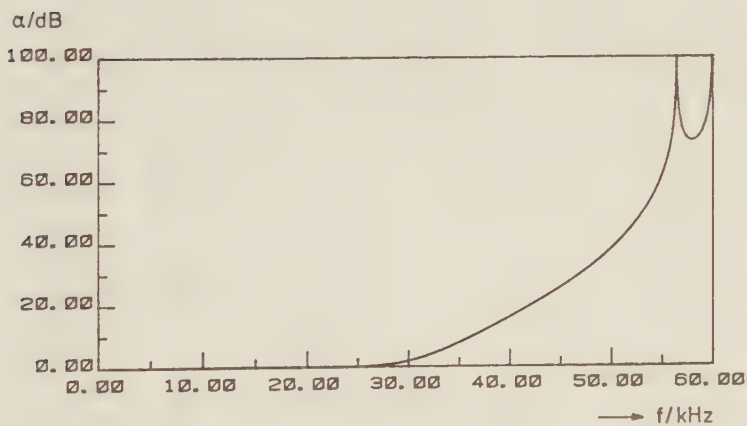
$$\gamma_1 = 0.1111 \cdot 2^{-1} \quad , \quad \gamma_2 = 0.101 \cdot 2^{-2} \quad .$$

Bei diesem Tiefpaß wurde dieselbe Untersuchung wie bei dem vorhergehenden Bandpaß mit der Ausnahme durchgeführt, daß die Störung durch ein starkes weißes Rauschen verursacht wurde, so daß wiederum auch im Innern des Filters Überläufe auftraten. Bild 4.20 zeigt die erhaltenen Signalverläufe des gestörten Eingangssignals, des zugehörigen Ausgangssignals $y(mT)$, des ungestörten Ausgangssignals $y_o(mT)$ und des Differenzsignals $e(mT)$.

Diese beiden Beispiele bestätigen somit die erwartete Antwortstabilität der Wellendigitalfilter.



(a)



(b)

Bild 4.19: a) Wellenflußdiagramm des Filters 3.Grades,
b) zugehöriger Dämpfungsverlauf.

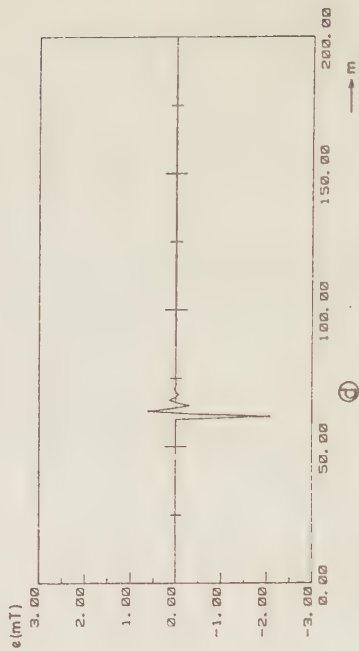
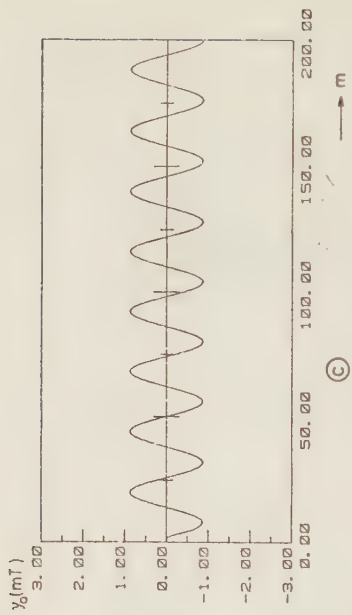
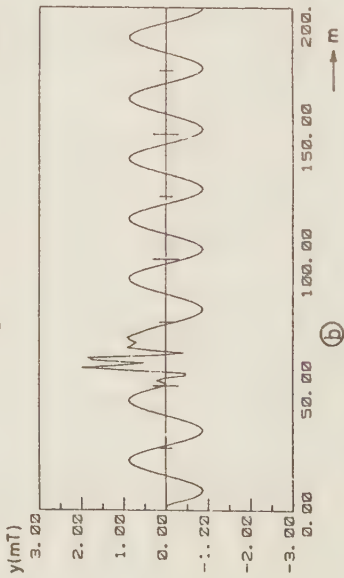
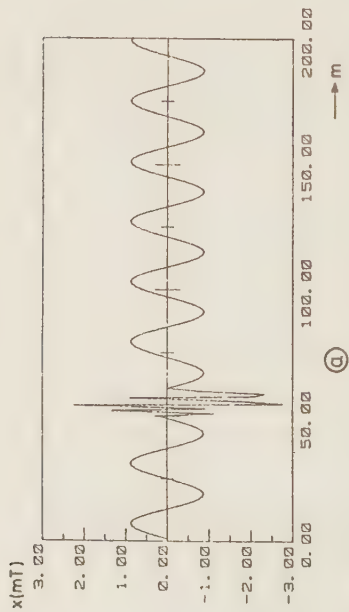


Bild 4.20: a) Gestörtes Eingangssignal, b) zugehöriges Ausgangssignal, c) ungestörtes Ausgangssignal, d) Differenzsignal $e(mT)$.

Schließlich werden wir das Ergebnis einer Simulation zur Untersuchung der Schleifenstabilität vorstellen. Es ist bekannt, daß die Hochpässe konventioneller Art unter Schleifenbedingungen besonders kritisch sind [66]. Deshalb betrachten wir einen Hochpaß 7. Grades, dessen Koeffizientenwerte wie folgt angegeben sind:

$$\gamma_1 = 0.1 \cdot 2^0, \quad \gamma_2 = 0.111 \cdot 2^0, \quad \gamma_3 = 0.1001 \cdot 2^{-2}$$

Bild 4.21 zeigt das Wellenflußdiagramm und den zugehörigen Dämpfungsverlauf.

Es wurden Ein- und Ausgang des Filters so verbunden, daß ein rückgekoppeltes System gemäß Bild 4.22 entstand. Die minimale Schleifenfäpfung betrug etwa 0.5 dB.

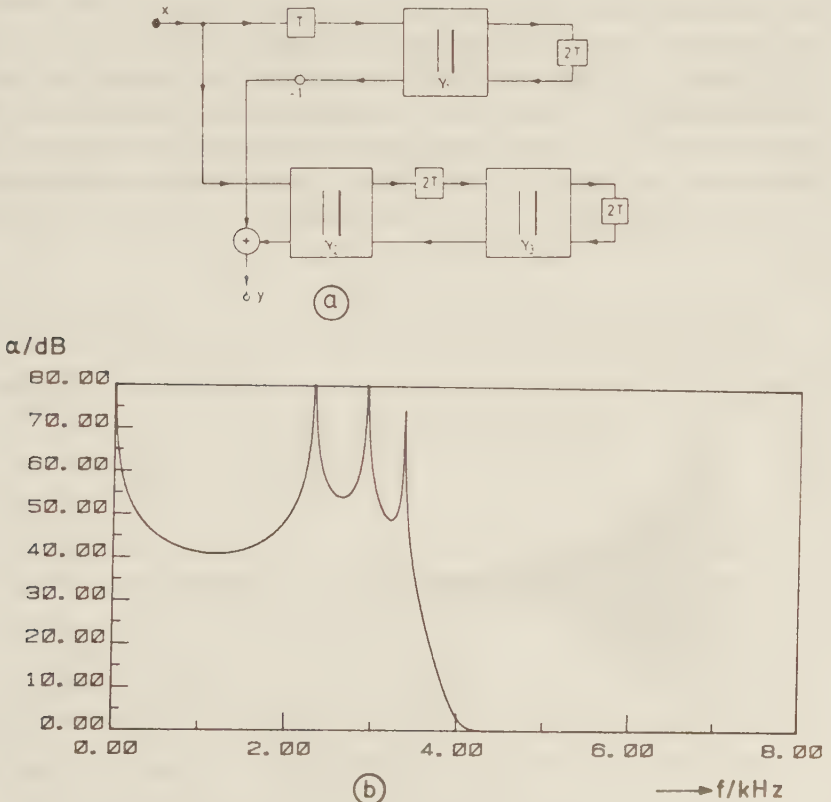


Bild 4.21: a) Wellenflußdiagramm des Hochpasses 7. Grades, zugehöriger Dämpfungsverlauf.
b)

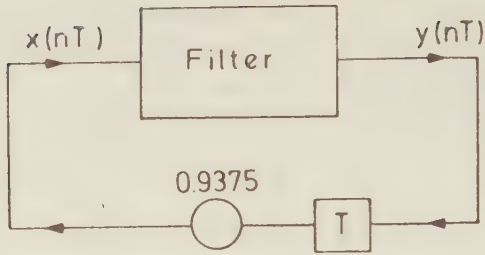


Bild 4.22: Wellendigitalfilter unter Schleifenbedingungen.

Es wurden dann als Anfangszustände, alle Speicherwerte auf den Wert -8 gesetzt. Darüberhinaus wurden die Signalparameter aus dem Gleitkommasystem $GS(2,10,-4,3)$ gewählt. Zusätzlich wurde Betrags-schneiden an allen Adaptoraustritten angewandt.

Bild 4.23 stellt den dafür erhaltenen zeitlichen Verlauf des Signals $y(mT)$ bzw. $x(mT)$. Wie zu erkennen ist, konvergiert die Signalamplitude von anfänglich großen Werten sehr rasch gegen Null. Aus dieser Betrachtung ergibt sich, daß keine Instabilitäten von Wellendigitalfiltern, die im Vierdrahtteil einer Nachrichtenverbin-dung eingesetzt sind, verursacht werden.

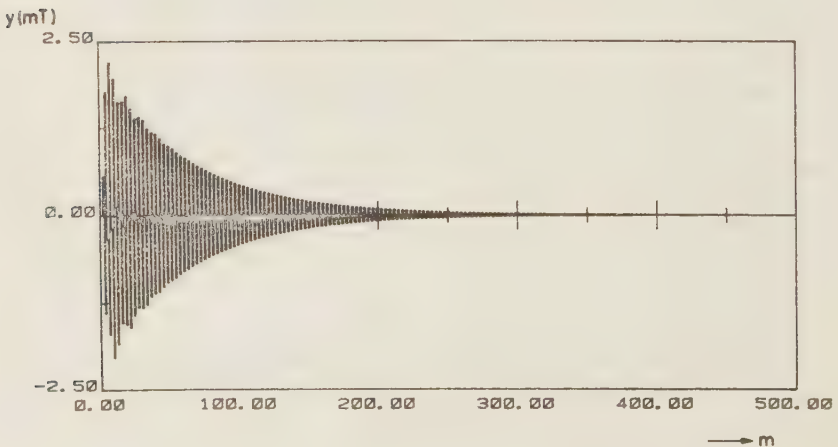


Bild 4.23: Ausgangssignal $y(mT)$.

5. Stabilitätsuntersuchungen von konventionellen rekursiven Digitalfiltern 2. Ordnung

Wir haben in Kapitel 4 festgestellt, daß im Fall einer Zweierkomplement-Gleitkommaarithmetik Wellendigitalfilter völlig ohne Instabilitätsprobleme realisiert werden können. In diesem Kapitel werden wir Stabilitätsuntersuchungen an konventionellen Digitalfiltern durchführen, um zu zeigen, daß Grenzzyklusschwingungen in solchen Filtern sowohl bei autonomen (Eingangssignal = 0) als auch bei rückgekoppelten Betriebsbedingungen entstehen können.

Bekanntlich lassen sich Digitalfilter konventioneller Art als Kaskadenschaltung von Gliedern erster und zweiter Ordnung realisieren, um mit einer kürzeren Wortlänge für die Multipliziererkoeffizienten auskommen zu können. Man kann sich leicht davon überzeugen, daß die Stabilität der gesamten Filteranordnung in direkter Weise von der Stabilität der einzelnen Glieder der Kaskade abhängt, und daß ein Glied erster Ordnung ein Sonderfall eines Gliedes zweiter Ordnung ist. Aus diesen Gründen werden wir uns im folgenden damit beschäftigen, inwieweit die Stabilität eines Gliedes zweiter Ordnung unter den oben genannten Betriebsbedingungen beeinträchtigt wird.

5.1 Grenzzyklusschwingung bei abgeschaltetem Eingangssignal

Die Übertragungsfunktion eines zweigradigen Gliedes konventioneller Art wird im allgemeinen als

$$H(z) = \frac{A_0 z^2 + A_1 z + A_2}{z^2 + B_1 z + B_2} \quad (5.1)$$

angegeben. Hierin sind A_0, A_1, A_2, B_1 und B_2 reelle Konstanten. Bild 5.1 stellt die kanonische Realisierung dieser Übertragungsfunktion dar.

Die Stabilität eines rekursiven Filters hängt letztlich nur von den Werten der Nennerkoeffizienten ab. Sie müssen die nachstehenden Bedingungen erfüllen:

$$|B_2| < 1 \quad \text{und} \quad |B_1| < B_2 + 1 \quad . \quad (5.2)$$

Diese Ungleichungen beschreiben den in Bild 5.2 dargestellten bekannten dreieckförmigen Stabilitätsbereich.

Wie bereits in Kapitel 4 erwähnt, sind rekursive Digitalfilter wegen der Wortlängenbegrenzung keine lineare Systeme. Hierdurch können parasitäre Schwingungen entstehen, auch wenn das Filter unter idealen linearen Bedingungen stabil ist. Bevor wir in die Analyse dieses Problems eingehen, nehmen wir an, daß die beiden Multiplikationen im rekursiven Teil exakt durchgeführt werden und keine Exponentenüberläufe auftreten.

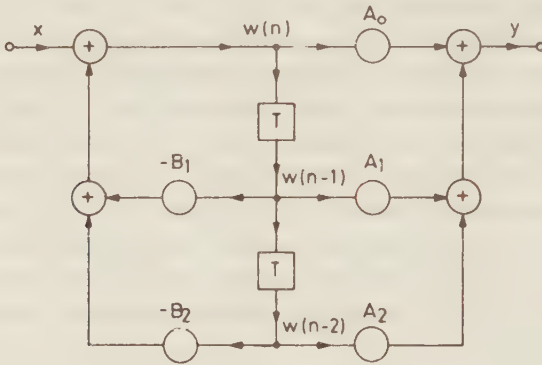


Bild 5.1: Kanonische Realisierung eines Digitalfilters 2. Ordnung.

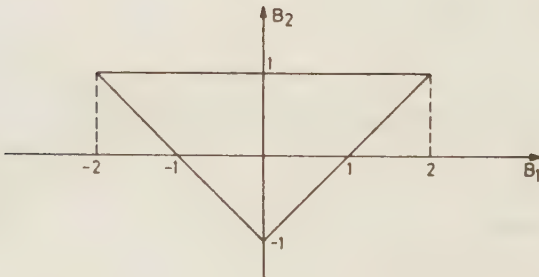


Bild 5.2: Stabilitätsbereich der Nennerkoeffizienten B_1 und B_2 .

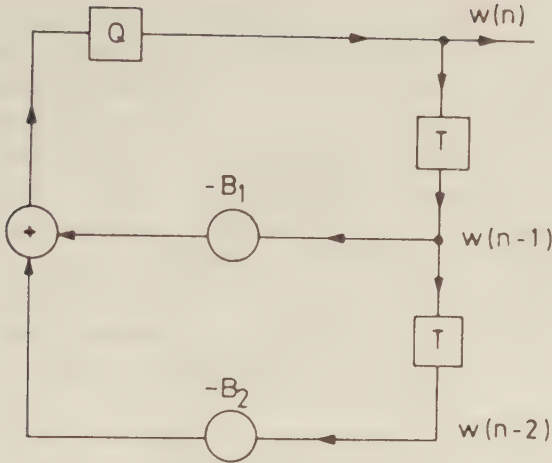


Bild 5.3: Rekursiver Teil eines Digitalfilters 2. Ordnung.

Da der nichtrekursive Teil keinen Einfluß auf die Stabilität des Systems ausübt, wird sich unser Interesse besonders auf die Stabilitätseigenschaften des rekursiven Teils konzentrieren. Bild 5.3 zeigt den rekursiven Teil, dessen Differenzengleichung lautet:

$$w(n) = Q[-B_1 w(n-1) - B_2 w(n-2)] \quad , \quad (5.3)$$

wobei Q einen Gleitkommaquantisierer bezeichnet und $n=mT$ ($m=-1, 0, 1, 2, \dots$) ist.

Betrachten wir den Fall des Betragsschneidens, da das Runden besonders kritisch ist, so erhalten wir

$$w(n) + B_1 w(n-1) + B_2 w(n-2) = e(n) \quad , \quad (5.4)$$

mit

$$e(n) = [B_1 w(n-1) + B_2 w(n-2)] \varepsilon_r(n) \quad , \quad (5.5)$$

wobei ε_r der relative Fehler ist.

Es ist bekannt, daß eine periodische Lösung mit der Periode L , nämlich eine Lösung mit den Elementen $w(n+1), w(n+2) \dots w(n+L)$ und mit den zugehörigen Fehlern $e(n+1), e(n+2) \dots e(n+L)$, die Gleichung (5.4) erfüllen kann [38, 63]. Das bedeutet, daß das Erhalten von Grenzzyklusschwingungen für bestimmte Werte von B_1 und B_2 nicht aus-

geschlossen ist. Solche Schwingungen können sogar den maximal darstellbaren Exponenten besitzen. Um dies nachzuweisen, betrachten wir die Differenzengleichung (5.4). Da die Multiplikation mit 2^e den Wert des relativen Fehlers ε_r (Gl. (5.5)) keine Änderung verursacht, kann man sich davon überzeugen, daß eine periodische Lösung für (5.4) mit den Elementen $w(n+1) 2^e, \dots, w(n+L) 2^e$ und $e(n+1) 2^e, \dots, e(n+L) 2^e$ entstehen kann [47]. Selbstverständlich kann $e=e_{\max}$ sein. In diesem Fall ist das digitale System unbrauchbar.

Zur Demonstration dieses Vorgangs betrachten wir als Beispiel den rekursiven Teil eines konventionellen Digitalfilters 2. Ordnung (Bild 5.3). Die zugehörigen Koeffizientenwerte sind wie folgt angegeben:

$$B_1 = -1.83984375 = -0.111010111 \cdot 2^1, \quad ,$$

$$B_2 = 0.87109375 = 0.11011111 \cdot 2^0 \quad .$$

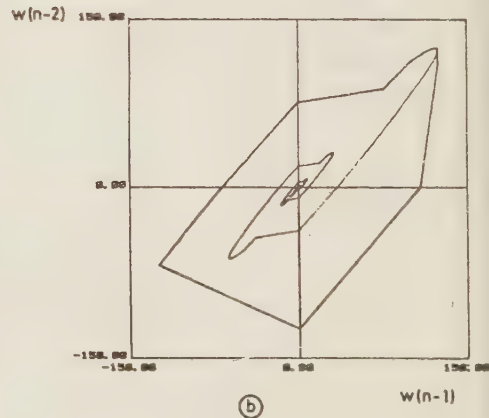
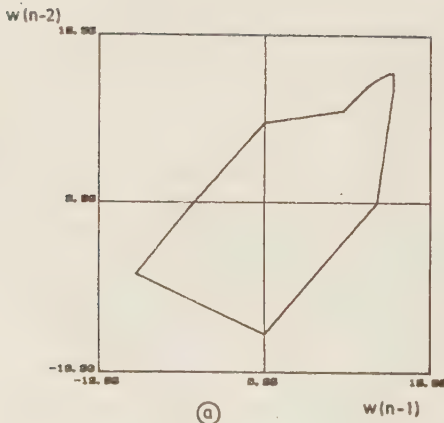


Bild 5.4: a) Grenzzyklus mit dem Gleitkommasystem $GS(2,6,-8,3)$,
b) Grenzzyklus mit dem Gleitkommasystem $GS(2,6,-8,7)$.

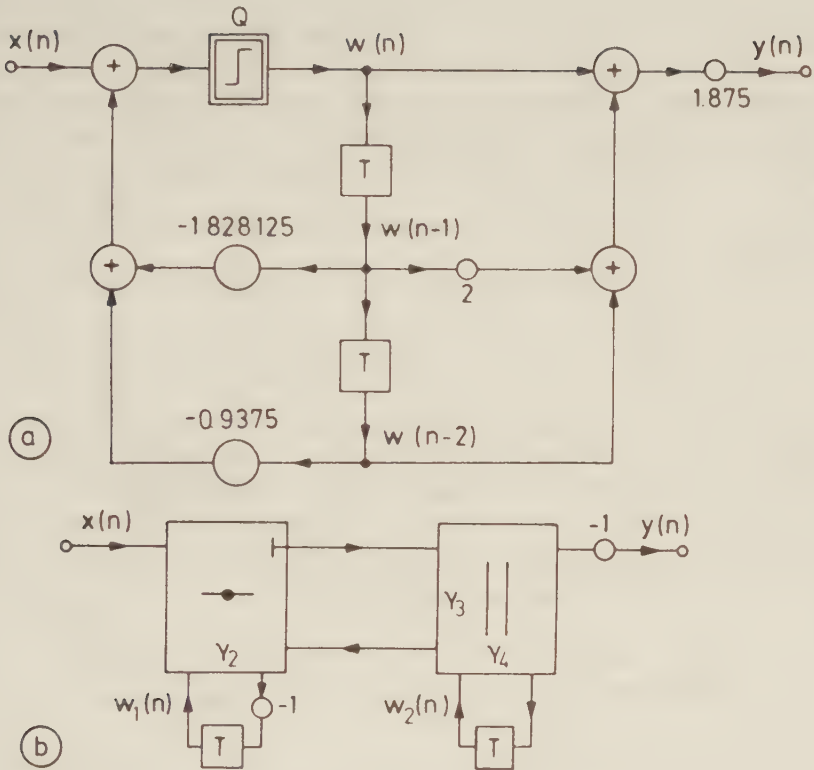


Bild 5.5: a) Ein Digitalfilter 2. Ordnung.
b) Ein Wellendigitalfilter 2. Ordnung.

Es wurden die Signalparameter aus dem Gleitkommasystem $GS(2,6,-8,3)$ und als Anfangszustände, die Speicherwerte jeweils -8 gewählt. Dann wurde Betragsschneiden mit Sättigungskennlinie nach der Addition angewandt. Bild 5.4a zeigt den dafür erhaltenen Grenzzyklus.

Dann wurde der maximale Exponent auf 7 gesetzt, d.h., daß die Signalparameter aus dem Gleitkommasystem $GS(2,6,-8,7)$ gewählt wurden. Bild 5.4b zeigt den dafür erhaltenen Grenzzyklus. Wie aus Bild 5.4 zu erkennen ist, entstanden Grenzzyklen mit großer Amplitude, die den maximal darstellbaren Exponenten erreicht.

Kapitel 4 hat gezeigt, daß sich Wellendigitalfilter im Gegensatz zu den konventionellen rekursiven Digitalfiltern völlig frei von Grenzzyklusschwingungen realisieren lassen, obwohl es sich bei solchen Filtern um stark vermascht rückgekoppelte Systeme handelt. Zum Ver-

gleich zwischen beiden Filtern, im Hinblick auf die Stabilitätseigenschaften bei abgeschaltetem Eingangssignal ($x(n)=0$), betrachten wir zwei Filterbeispiele, einerseits ein konventionelles rekursives Digitalfilter 2. Ordnung (vgl. Bild 5.5a) und andererseits ein Wellendigitalfilter 2. Ordnung (vgl. Bild 5.5b). Wählt man die Koeffizientenwerte des Wellendigitalfilters wie folgt:

$$\gamma_2 = 0.111 \cdot 2^0, \quad \gamma_3 = 0.1111 \cdot 2^1, \quad \gamma_4 = 0.1 \cdot 2^{-3},$$

so erhält man für die beiden Filter dieselbe Übertragungsfunktion

$$H(z) = \frac{Y(z)}{X(z)} = 1.875 \frac{z^2 + 2z + 1}{z^2 + 1.828125z + 0.9375}$$

Es wurden die Signalparameter aus dem Gleitkommasystem GS(2,6,-4,3) gewählt, und als Anfangszustände wurden alle Speicherwerte auf den Wert -8 gesetzt. Diese Aussage gilt für die beiden Filter.

Wir wenden uns zunächst dem konventionellen Digitalfilter zu. Es wurde Betragsschneiden nach der Addition angewandt. Bild 5.6 zeigt den dafür erhaltenen Grenzyklus.

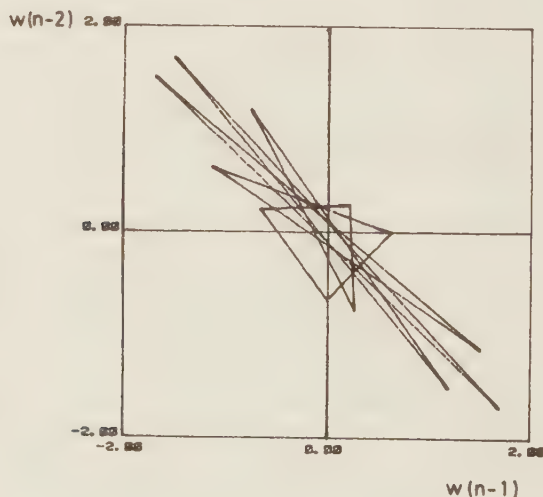


Bild 5.6: Grenzyklus im Filter gemäß Bild 5.5a.

Als nächstes betrachten wir das Wellendigitalfilter, dessen Wellenflußdiagramm in Bild 5.5b dargestellt ist. Die im Kapitel 4

diskutierten Maßnahmen, die die Pseudopassivität eines Wellendigitalfilters gewährleisten, wurden an allen Adaptoraustritten angewandt. Bild 5.7 stellt den dafür erhaltenen zeitlichen Verlauf der in Bild 5.5b eingezeichneten Signale $w_1(n)$ bzw. $w_2(n)$ dar. Es ist ersichtlich, daß die Signalamplituden von anfangs großen Werten sehr rasch abnehmen. Tatsächlich wurde

$$w_1(n) = w_2(n) = 0 \quad \text{für } m > 300$$

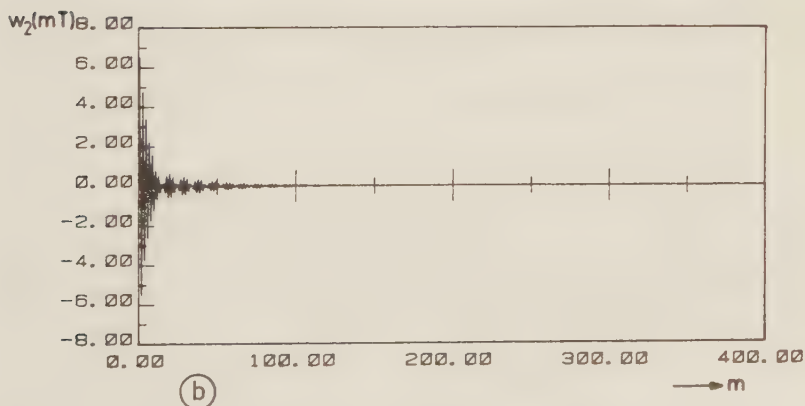
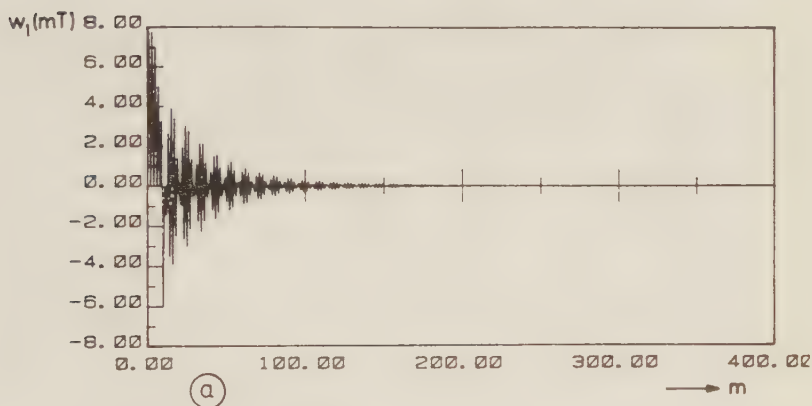


Bild 5.7: a) Das Signal $w_1(n)$, b) das Signal $w_2(n)$.

erreicht. Dieses Beispiel bestätigt die ausgezeichneten Stabilitätseigenschaften von Wellendigitalfiltern. Aus diesem Grunde muß der rekursive Teil in einem konventionellen Digitalfilter 2. Ordnung

als Wellendigitalfilter realisiert werden [64,65].

5.2 Ein auf Wellendigitalfilter beruhendes Gleitkomma-Digitalfilter 2. Ordnung ohne Grenzzyklusschwingung

Ausgehend von der Struktur eines einfachen Wellendigitalfilters 2. Ordnung (vgl. Bild 5.8) erhalten wir bei abgeschaltetem Eingangssignal

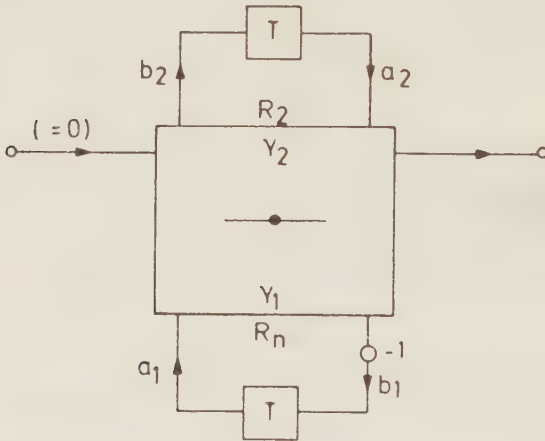


Bild 5.8: Ein Wellendigitalfilter 2. Ordnung.

$$b_1 = (\gamma_1 - 1)a_1 + \gamma_1 a_2 \quad , \quad (5.6a)$$

$$b_2 = -\gamma_2 a_1 + (1 - \gamma_2) a_2 \quad . \quad (5.6b)$$

Durch geeignete Anzapfung des Eingangssignals x kann ein allgemeines Glied 2. Ordnung realisiert werden. Bild 5.9 stellt die sich ergebende Schaltung dar. Die zugehörige Übertragungsfunktion läßt sich folgendermaßen schreiben:

$$H(z) = \frac{z^2 + A_1 z + A_2}{z^2 + B_1 z + B_2} \quad (5.7a)$$

$$\begin{aligned} \text{mit } A_1 &= -(\beta_1 + \beta_2) \quad , \quad A_2 = 1 + \beta_1 - \beta_2 \quad , \\ B_1 &= \gamma_2 - \gamma_1 \quad , \quad B_2 = \gamma_1 + \gamma_2 - 1 \quad . \end{aligned} \quad (5.7b)$$

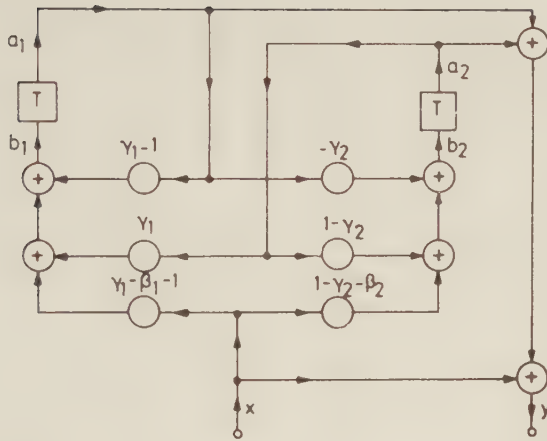


Bild 5.9: Ein Digitalfilter 2. Ordnung ohne Grenzzyklen.

Die sich aus (5.7b) ergebenden Multipliziererkoeffizienten (vgl. Bild 5.9) sind wie folgt angegeben:

$$\begin{aligned} \gamma_1 &= \frac{1+B_2-B_1}{2} \quad , \quad \gamma_1-1 = \frac{B_2-B_1-1}{2} \quad , \\ \gamma_2 &= \frac{1+B_1+B_2}{2} \quad , \quad 1-\gamma_2 = \frac{1-B_1-B_2}{2} \quad , \\ \gamma_1-\beta_1-1 &= \frac{B_2+A_1-B_1-A_2}{2} \quad , \quad 1-\gamma_2-\beta_2 = \frac{A_1+A_2-B_1-B_2}{2} \quad . \end{aligned} \quad (5.8)$$

Wie zu erkennen ist, lassen sich die Multipliziererkoeffizienten in Abhängigkeit der Koeffizienten der Übertragungsfunktion einfach ohne Einschränkung realisieren. Um dies zu erklären, betrachten wir die Stabilitätsbedingungen (5.2). Wir erhalten die folgenden Beziehungen

$$\gamma_1 + \gamma_2 < 2 \quad , \quad \gamma_1 > 0 \quad , \quad \gamma_2 > 0 \quad . \quad (5.9)$$

Diese Bedingungen sind automatisch erfüllt, da γ_1 und γ_2 die Adaptorkoeffizienten des in Bild 5.8 dargestellten Wellendigitalfilters bezeichnen.

Um die Pseudopassivität des rekursiven Teils zu erhalten, muß die Bedingung (4.5) erfüllt werden. Mit Hilfe der in Kapitel 4 diskutierten Maßnahmen bei Wellendigitalfiltern kann dieses Glied, dessen Konfiguration in Bild 5.9 gezeichnet ist, garantiert frei von Grenzyklusschwingungen und auch antwortstabil realisiert werden. Ein solches Glied ist noch stets kanonisch bezüglich der Anzahl der Verzögerer. Die Anzahl der Addierer und der Multiplizierer wird jeweils von 4 auf 6 erhöht.

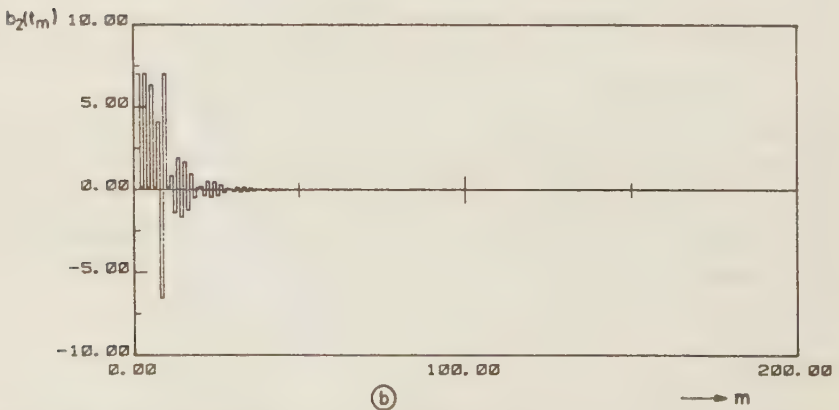
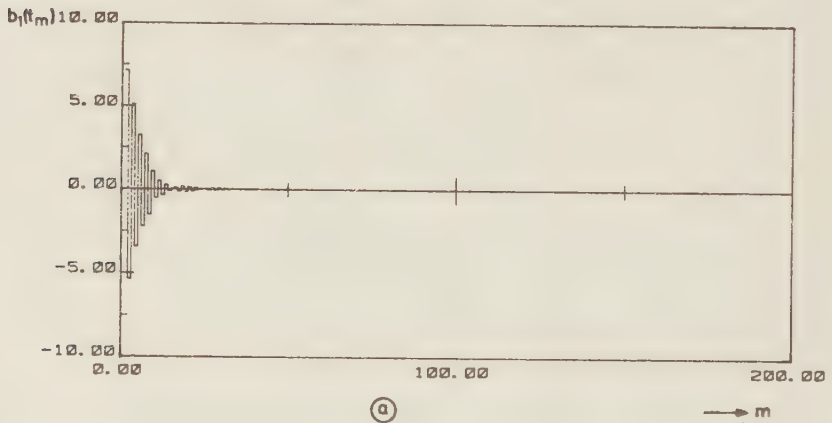


Bild 5.10: a) Signal b_1 , b) Signal b_2 .

Es soll nun noch das Ergebnis einer Simulation zur Erörterung der Stabilität dieses Gliedes bei abgeschaltetem Eingangssignal vorgestellt werden. Um dieselben Nennerkoeffizienten wie bei dem Beispiel 5.5a zu erhalten, wählen wir

$$\gamma_1 = 0.0546875 = 0.0000111 \quad , \quad \gamma_2 = 1.8828125 = 1.1110001 \quad .$$

Es wurden dieselben Maßnahmen zur Erhaltung der Pseudopassivität der Wellendigitalfilter nach der Addition angewandt. Die Signalparameter wurden aus dem Gleitkommasystem GS(2,3,-4,3) gewählt (also mit einer Mantisse, die nur 3 Bit umfaßt), um möglichst ungünstige Betriebsbedingungen zu simulieren. Als Anfangszustände wurden alle Speicherwerte auf -8 gesetzt.

Bild 5.10 zeigt den dafür erhaltenen zeitlichen Verlauf des Signals b_1 bzw. b_2 (vgl. Bild 5.9) für einen Ausschnitt von 200 Abtastwerten. Wie zu erkennen ist, konvergieren die Signalamplituden sehr rasch gegen Null. Dieses Beispiel bestätigt somit das erwartete Stabilitätsverhalten.

5.3 Stabilität der Nachrichtenschleife

In [50] wurde nachgewiesen, daß Wellendigitalfilter unter Schleifenbedingungen völlig frei von Grenzyklusschwingungen realisiert werden, wenn die Pseudopassivität solcher Filter garantiert wird. Es ist lediglich erforderlich, daß die in Kapitel 4 beschriebenen Maßnahmen zur Unterdrückung von Grenzyklusschwingungen durchgeführt werden.

Die Anforderungen an die Stabilität von konventionellen Digitalfiltern, die sich im Vierdrahtteil einer Nachrichtenverbindung befinden, sind strenger als an diejenigen, die unter nichtrückgekoppelten Betriebsbedingungen arbeiten. Der Grund besteht darin, daß unter ungünstigen Umständen wegen der nichtkontrollierbaren Phasendrehung parasitäre Schwingungen sogar mit großer Amplitude auftreten können, auch wenn die Digitalfilter im isolierten Fall völlig grenzzyklusfrei sind.

Bisher ist unbekannt, unter welchen Bedingungen konventionelle rekursive Digitalfilter in Kaskadenstruktur, die in Festkommaarithmetik realisiert sind, Schleifenstabilität besitzen können [1,66]. Im folgenden werden wir einige Untersuchungen durchführen, um festzustellen, ob die konventionellen rekursiven Gleitkomma-Digitalfil-

ter die Schleifenstabilität besitzen können, bzw. wie diese Stabilität beeinträchtigt wird. Da die Stabilität der gesamten Filteranordnung in direkter Weise von der Stabilität der einzelnen Glieder der Kaskade abhängt, werden sich die Untersuchungen auf Glieder 2. Ordnung, die im isolierten Fall völlig frei von Grenzyklen sind, unter rückgekoppelten Betriebsbedingungen konzentrieren. Allerdings würde die Situation noch weitaus kritischer, wenn das Filter aus mehreren Gliedern 2. Ordnung besteht, wie dies in der Praxis der Fall ist.

Für die Untersuchungen des Stabilitätsverhaltens eines Digitalfilters unter Schleifenbedingungen wurden Simulationen in einer Anordnung gemäß Bild 5.11 durchgeführt. Der Block F stellt das zu untersuchende Digitalfilter dar. Durch den Block S wird F rückgekoppelt.

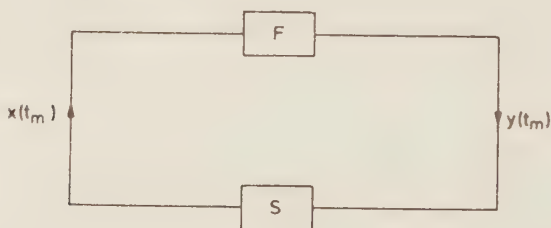


Bild 5.11: Untersuchungsanordnung.

Es wurde ein konventionelles rekursives Digitalfilter 2. Ordnung gewählt, da sich, wie schon erwähnt wurde, Digitalfilter höherer Ordnung als Kaskadenstruktur von Gliedern 1. und 2. Ordnung realisieren lassen. Dann wurde die Anordnung gemäß Bild 5.11 untersucht. Der Rückkopplungspfad besteht aus 8 Verzögerungsgliedern in Reihe mit einem Multiplizierer γ_0 (vgl. Bild 5.12). Die minimale Schleifendämpfung bei einer Frequenz, die im Durchlaßbereich des Filters liegt, betrug etwa 6.2 dB. Die arithmetischen Operationen, nämlich die Multiplikationen und die Additionen, wurden mit der vollen Genauigkeit des Rechners HP1000 ausgeführt, d.h., die Ergebnisse der beiden Multiplikationen wurden in voller Länge addiert. Mit dem Quantisierer Q (Bild 5.12a) wurden die Signalparameter aus dem Gleitkommasystem GS(2,6,-4,3) durch Betragsschneiden gewählt.

Als Anfangszustand wurden alle Speicherwerte auf -8 gesetzt.

konventionellen Filters.

Parallel hierzu wurde ein Wellendigitalfilter 2. Ordnung mit derselben Übertragungsfunktion und mit denselben Anfangszuständen wie bei dem konventionellen Digitalfilter unter denselben Schleifenbedingungen untersucht. Bild 5.13b stellt die erhaltenen zeitlichen Verläufe der Zustandsvariablen im Wellendigitalfilter dar. Darüber hinaus wurde die Schleifendämpfung etwa 0.2 dB ($\gamma_0 \approx -0.046875$) gewählt, trotzdem traten keine Schwingungen innerhalb des Wellendigitalfilters auf. Bild 5.14 zeigt den erhaltenen zeitlichen Verlauf der Zustandsvariablen. Dieses Beispiel bestätigt somit das erwartete Stabilitätsverhalten der Wellendigitalfilter unter Schleifenbedingung und stellt fest, daß konventionelle rekursive Digitalfilter 2. Ordnung unter Schleifenbedingungen nicht frei von Instabilitäten sind, auch wenn die Schleifendämpfung positiv ist.

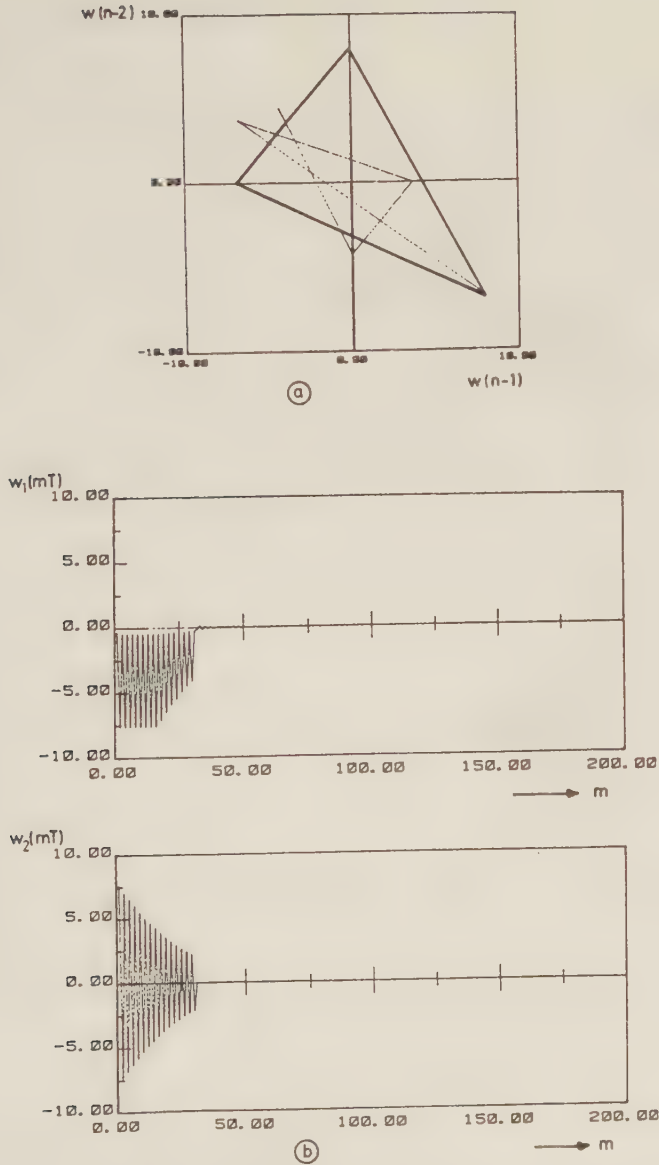


Bild 5.13: Verlauf der Zustandsvariablen unter Schleifenbedingung innerhalb des a) konventionellen Digitalfilters, b) Wellendigitalfilters. (Schleifendämpfung etwa 6,2 dB).

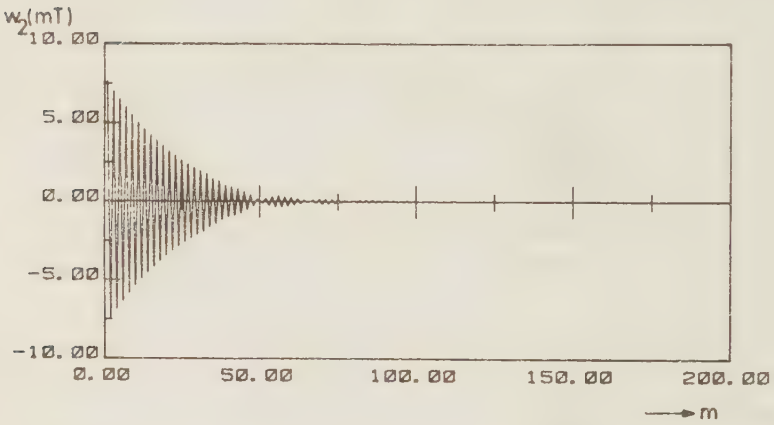
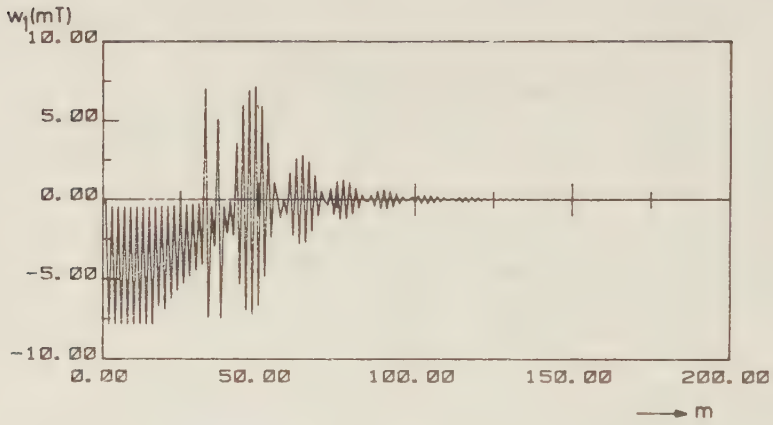


Bild 5.14: Verlauf der Zustandsvariablen im Wellendigitalfilter (Schleifendämpfung etwa 0,2 dB).

6. Berechnung des Signal-Rausch-Verhältnisses bei Gleitkommaarithmetik

6.1 Vorbemerkungen

Aufgrund des notwendigerweise begrenzten Hardwarerealisierungsaufwandes können in einem "realen" Digitalfilter nur endlich viele verschiedene Zahlen dargestellt und verarbeitet werden. Die Begrenzung der Multipliziererkoefizientenwortlänge führt zur Abweichung des Dämpfungsverlaufs eines Digitalfilters von dem ursprünglich angestrebten Verlauf. In diesem Fall spricht man von dem sogenannten "Einstellfehler". Die Lösung dieses Problems besteht darin, daß man ausgehend von den idealen Multipliziererkoefizienten durch diskrete Optimierung eine erforderliche Wortlänge zur Erfüllung des Toleranzschemas bestimmen kann. Aus diesem Grunde sind Digitalfilterstrukturen, deren Empfindlichkeit der Dämpfung gegenüber Änderungen der Multipliziererkoefizienten klein ist, von besonderer Bedeutung. Es hat sich gezeigt, daß Wellendigitalfilter häufig mit kurzer Koefizientenwortlänge zu realisieren sind [16-19].

Die Darstellung der Signalparameter mit endlicher Wortlänge führt in einem Digitalfilter zu Fehlern, die als "Betriebsfehler" bezeichnet werden. Diese lassen sich in rauschartige Fehler und Grenzyklusschwingungen klassifizieren. Rauschartige Fehler, die das Signal-Rausch-Verhältnis am Ausgang des Digitalfilters herabsetzen, haben ihre Ursache in den notwendigen Rundungs- und/oder Schneideoperationen der Mantisse des Ergebnisses nach bestimmten arithmetischen Operationen. Eine weitere Ursache liegt in den Exponentenüberläufen. Wie schon im Kapitel 4 diskutiert wurde, können Gleitkomma-Wellendigitalfilter völlig frei von Grenzyklusschwingungen realisiert werden.

In diesem Kapitel befassen wir uns mit der Analyse des Signal-Rausch-Verhältnisses, das den Abstand, üblicherweise in dB ausgedrückt, zwischen Nutzsignalleistung und Rauschleistung angibt. Bevor wir auf die Analyse eingehen, werden wir einige Begriffe aus der Theorie der digitalen Signalverarbeitung vorstellen.

Ein zeitdiskretes Signal, $x(t_m)$, läßt sich durch äquidistante Abtastung von einem zeitkontinuierlichen Signal $x(t)$ zu den Zeitpunkten t_m ($t_m = t_0 + mT$, $m = \dots, -1, 0, 1, 2, \dots$) gewinnen, und hat die mittlere Leistung, die durch

$$P_x = \lim_{N \rightarrow \infty} \frac{1}{(2N+1)T} \sum_{m=-N}^N |x(t_m)|^2 \quad (6.1)$$

definiert ist. Durch Mittelung über einen bestimmten Zeitraum ergibt sich daraus

$$\hat{P}_x = \frac{1}{(2N+1)T} \sum_{m=-N}^N |x(t_m)|^2 \quad (6.2)$$

Dabei stellt $(2N+1)T$ das Mittelungsintervall dar.

Zur Berechnung des Signal-Rausch-Verhältnisses SNR (nach dem englischen Signal to Noise Ratio) wird von der Beziehung

$$SNR = 10 \log \frac{\hat{P}_S}{\hat{P}_N} \quad (6.3)$$

Gebrauch gemacht, wobei $\hat{P}_S$ die mittlere Leistung des Nutzsignals und $\hat{P}_N$ die mittlere Leistung des Störsignals bezeichnet. Nimmt man an, daß Nutz- und Störsignale rauschartig durch ergodische stationäre und mittelwertfreie Zufallsprozesse erzeugt werden, so kann die Berechnung des SNR auch über die Varianzen der beiden Signale erfolgen.

$$SNR = 10 \log \frac{\sigma_S^2}{\sigma_N^2} \quad (6.4)$$

Für die Untersuchung der Quantisierungseffekte innerhalb eines Digitalfilters ist es zweckmäßig, einen einzelnen Quantisierer zu untersuchen.

6.2 Berechnung des Rauschens eines Quantisierers

Da für die Berechnung des Quantisierungsrauschens eines Quantisierers der absolute Fehler ϵ_{abs} Verwendung findet, gilt das Ersatzbild nach 6.1b mit der additiven Fehlereinwirkung

Wie bei Festkommaarithmetik kann der Einfluß des Gleitkomma-Quantisierungsfehlers durch eine additive Rauschquelle berücksichtigt werden. In diesem Fall ist jedoch die Leistung der Rauschquelle vom Wert des arithmetischen Ergebnisses abhängig.

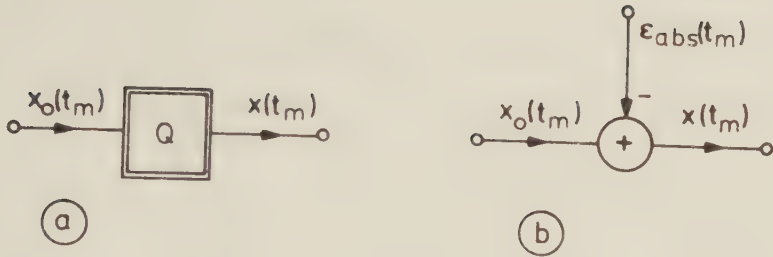


Bild 6.1: a) Quantisierung des Signals $x_o(t_m)$ mit dem Quantisierer Q ,
b) Ersatzbild des Quantisierers nach a) .

6.2.1 Rundungsrauschen

Wir haben im Kapitel 3 gesehen, daß sich das Signal x (vgl. (3.8)) und der absolute Fehler ϵ_{abs} (vgl. (3.24)) wie folgt schreiben lassen:

$$x = m \cdot 2^e = m \cdot y \quad , \quad \epsilon_{abs} = \epsilon \cdot 2^e = \epsilon \cdot y \quad ,$$

solange keine Exponentenüberläufe und keine Unterläufe auftreten. Bei Rundungsoperationen kann man annehmen, daß keine statistische Abhängigkeit zwischen x und ϵ_{abs} besteht. Dann erhalten wir im Falle stationärer ergodischer Signale unter der Voraussetzung der statistischen Unabhängigkeit zwischen m und y sowie zwischen ϵ und y und der weiteren Voraussetzung, daß keine Fehler durch Begrenzung des Zahlenbereichs auftreten

$$\begin{aligned} \sigma_x^2 &= E[x^2] - E^2[x] = E[x^2] = E[m^2] \cdot E[y^2] \quad \text{da } E[x] = 0 \quad , \\ \sigma_{\epsilon_{abs}}^2 &= E[\epsilon_{abs}^2] - E^2[\epsilon_{abs}] = E[\epsilon_{abs}^2] = E[\epsilon^2] \cdot E[y^2] \quad \text{da } E[\epsilon_{abs}] = 0 \\ E[m^2] &= \sigma_m^2 \quad \text{da } E[m] = 0 \quad , \\ E[\epsilon^2] &= \sigma_\epsilon^2 \quad \text{da } E[\epsilon] = 0 \quad . \end{aligned} \quad (6.5)$$

Das SNR läßt sich unter Verwendung von (6.5) schreiben als

$$SNR = 10 \log \frac{\sigma_x^2}{\sigma_{\epsilon_{abs}}^2} = 10 \log \frac{\sigma_m^2 \sigma_y^2}{\sigma_\epsilon^2 \sigma_y^2} = 10 \log \frac{\sigma_m^2}{\sigma_\epsilon^2} \quad \text{dB} \quad . \quad (6.6)$$

Unter der Annahme, daß ϵ in dem durch (3.25) festgelegten Intervall

gleichverteilt ist, erhält man

$$\sigma_{\epsilon}^2 = \frac{2^{-2t_{\text{man}}}}{12} \quad (6.7)$$

Setzt man (3.13) und (6.7) in (6.6) ein, so erhält man

$$\text{SNR} \approx 8,45 + 6,02 t_{\text{man}} \text{ dB} \quad (6.8)$$

Der Beziehung (6.8) ist zu entnehmen, daß das SNR für den einzelnen Quantisierer mit Gleitkommarundung einen konstanten Verlauf über der Signalleistung besitzt und nur von der Mantissenwortlänge abhängt.

Unter der Vernachlässigung der statistischen Abhängigkeit zwischen dem relativen Fehler ϵ_r und dem Signal x erhalten wir

$$\text{SNR} = 10 \log \frac{\sigma_x^2}{\sigma_{\epsilon_r}^2 \sigma_x^2} = 10 \log \frac{1}{\sigma_{\epsilon_r}^2} \quad (6.9)$$

Für Gleitkommarundung entnimmt man der Tabelle 3.3 die Varianz des relativen Fehlers und erhält

$$\text{SNR} \approx 7,78 + 6,02 t_{\text{man}} \text{ dB} \quad (6.10)$$

Aus den Beziehungen (6.8) und (6.10) ist ersichtlich, daß die Differenz zwischen den beiden Berechnungsmethoden vernachlässigbar ist. Es bleibt anzumerken, daß die Begrenzung des Zahlenbereiches den konstanten Verlauf des SNR's bezüglich der Signalleistung einschränkt. Wie schon in Kapitel 3 erwähnt wurde, kann man durch geeignete Wahl der Exponentenwortlänge den Zahlenbereich soweit ausdehnen, daß die Auswirkungen der Zahlenbereichsbegrenzung nicht stören. Daraus resultiert bei Gleitkommadarstellung ein überlegenes SNR, besonders bei betragsmäßig kleinen Signalen.

6.2.2 Rauschen bei Betragsschneiden

Beim Schneiden der Beträge ist die Annahme, daß keine statistische Abhängigkeit zwischen dem absoluten Fehler ϵ_{abs} und dem Signal x_o besteht, nicht ohne weiteres erlaubt, da zumindest das Vorzeichen des absoluten Fehlers direkt vom Vorzeichen des Signals abhängig ist (vgl. Bild 3.6b). Man kann sich leicht davon überzeugen, daß sich der absolute Fehler ϵ_b bei Betragsschneiden wie folgt

schreiben läßt:

$$\epsilon_b(t_m) = \epsilon_a(t_m) + 2 \frac{e^{-t_{man}}}{2} \text{sign}[x_o(t_m)] \quad , \quad (6.11)$$

wobei $\epsilon_a(t_m)$ die gleiche Art statistischer Verteilung wie der absolute Rundungsfehler hat. Wie aus (6.11) zu ersehen ist, besteht der absolute Fehler bei Betragsschneiden einerseits aus einem ähnlichen Anteil wie beim Runden, zusätzlich jedoch aus einem Anteil gleich

$$\frac{e^{-t_{man}}}{2} \text{sign } x_o, \text{ der von } x_o \text{ statistisch abhängig ist. Es ist}$$

aber nicht richtig, den letzteren Anteil einfach als zusätzliches Rauschen aufzufassen. Um dies zu erklären, betrachten wir ein Sinussignal. Der absolute Fehler läßt sich als

$$\epsilon_b(t_m) = \epsilon_a(t_m) + 2 \frac{e^{-t_{man}}}{2} \text{sign} [\sin(\omega_o t_m + \theta)] \quad (6.12)$$

schreiben. Unter Verwendung der Fourierreihe

$$\text{sign}[\sin(\omega_o t_m + \theta)] = \frac{4}{\pi} \sum_{\mu=0}^{\infty} \frac{\sin[(2\mu+1)(\omega_o t_m + \theta)]}{2\mu+1} \quad (6.13)$$

läßt sich (6.12) schreiben als

$$\epsilon_b(t_m) = \frac{2}{\pi} \frac{e^{-t_{man}} + 1}{2} \sin(\omega_o t_m + \theta) + \epsilon_a(t_m) + \frac{2}{\pi} \frac{e^{-t_{man}} + 1}{2} \sum_{\mu=1}^{\infty} \frac{\sin[(2\mu+1)(\omega_o t_m + \theta)]}{2\mu+1} \quad . \quad (6.14)$$

Der Gleichung (6.14) kann man leicht entnehmen, daß der Signum-Term (in 6.12) einen zu dem Nutzsignal gehörenden Anteil enthält. Fettweis hat vorgeschlagen, diesen Anteil dem Nutzsignal zuzurechnen [67]. Bei Verwendung dieses Vorschlags können die beiden letzten in (6.14) auf der rechten Seite stehenden Terme als Rauschen betrachtet werden. Unter der Voraussetzung statistischer Unabhängigkeit von $\epsilon_a(t_m)$ und $\sin(m\omega_o T + \theta)$ berechnet sich die Rauschleistung des ersten Terms zu

$$\sigma_{\epsilon_a}^2 = \sigma_{\epsilon_{abs}}^2 = \sigma_{\epsilon}^2 \sigma_Y^2 (\epsilon_{abs} = \epsilon \cdot 2^e = \epsilon \cdot y \text{ vgl. (3.24)}) \quad .$$

Die Rauschleistung des zweiten Terms läßt sich zweckmäßig aus der Differenz der Leistung der Rechteckfolgefunktion

$e^{-t_{\text{man}}}$ 2^{-1} $\text{sign}[\sin(\omega_0 t_m + \theta)]$ und ihrer zum Nutzsignal gehörenden Grundwelle berechnen [68], was den Wert

$$\sigma_Y^2 \cdot 2^{-2t_{\text{man}}} \cdot 2^{-2} \left(1 - \frac{8}{\pi^2}\right)$$

ergibt. Dann erhalten wir die gesamte Rauschleistung durch

$$\hat{P}_N = \sigma_\epsilon^2 \sigma_Y^2 + \sigma_Y^2 \cdot 2^{-2t_{\text{man}}} \left(\frac{1}{4} - \frac{2}{\pi^2}\right) \quad (6.15)$$

Das Signal-Rausch-Verhältnis läßt sich unter Verwendung von (6.3), (6.5) ($P_S = E[x^2]$) und (6.15) schreiben als

$$\text{SNR}_b = 10 \log \frac{\sigma_m^2}{\sigma_\epsilon^2 + 2^{-2t_{\text{man}}} \left(\frac{1}{4} - \frac{2}{\pi^2}\right)} \text{ dB} \quad (6.16)$$

Unter Verwendung von (3.13) und (6.7) erhalten wir

$$\text{SNR}_b \approx 6,02 t_{\text{man}} + 6,5 \text{ dB} \quad (6.17)$$

Vernachlässigt man die statistische Abhängigkeit zwischen dem absoluten Fehler ϵ_b und dem Signal, so erhält man

$$\hat{P}_N = E[\epsilon_b^2] = E[\epsilon_r^2 x^2] = E[\epsilon_r^2] \cdot E[x^2] = [\sigma_{\epsilon_r}^2 + E^2[\epsilon_r]] E[x^2] \quad (6.18)$$

Das SNR_b läßt sich unter Verwendung von (6.18) und mit Hilfe der Tabelle 3.3 ausdrücken als

$$\begin{aligned} \text{SNR}_b &= 10 \log \frac{E[x^2]}{E[\sigma_b^2]} = 10 \log \frac{1}{\sigma_{\epsilon_r}^2 + E^2[\epsilon_r]} \\ &\approx 6,02 t_{\text{man}} + 1,76 \text{ dB} \quad (6.19) \end{aligned}$$

Aus (6.19) ist ersichtlich, daß man eine zusätzliche Rauschleistung von etwa 4,7 dB gegenüber dem Ergebnis des Betragsschneidens unter Voraussetzung statistischer Abhängigkeit vom absoluten Fehler und Signal erhält (vgl. Gl. 6.17).

6.3 Vergleich zur Festkommaarithmetik

Bei Festkommaarithmetik wird häufig die darstellbare Zahl x auf den Bereich

$$-1 \leq x < 1 \quad (6.20)$$

festgelegt. Durch Begrenzung der Signalwortlänge auf t Binärstellen (ohne Vorzeichenbit) ergibt sich ein Quantisierungsfehler, der von der Quantisierungsart abhängt. Unter der Annahme, daß der Fehler gleichverteilt ist, ergibt sich die Wahrscheinlichkeitsdichtefunktion $p(\epsilon)$ für den Fall des Rundens nach Bild 6.2a, für den Fall des Zweierkomplement-Schneidens nach Bild 6.2b und für den Fall des Betragschneidens bei positivem Signal nach Bild 6.2b und bei negativem Signal nach Bild 6.2c, wobei $Q_f = 2^{-t}$ die Festkommaquantisierungsstufe bezeichnet.

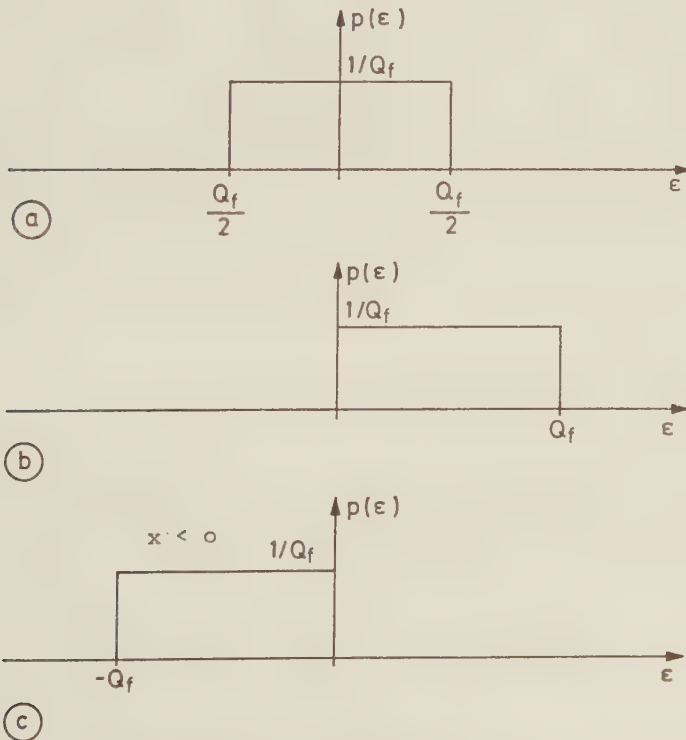


Bild 6.2: Wahrscheinlichkeitsdichtefunktion des Festkommaquantisierungsfehlers bei a) Runden, b) Zweierkomplement-schneiden und c) Betragschneiden.

Zunächst berechnen wir das SNR im Falle der Festkommarundung. Unter der Annahme, daß der Rundungsfehler gleichmäßig verteilt ist und daß keine Übersteuerung auftritt, erhält man

$$\hat{p}_N = \frac{2^{-2t}}{12} \quad (6.21)$$

Unter Verwendung von (6.3) und (6.21) läßt sich das SNR schreiben als

$$\text{SNR} \approx 6,02 t + 10,79 + 10 \log \hat{p}_s \quad \text{dB} \quad (6.22)$$

Wenn das Signal x in dem Bereich (6.20) gleichverteilt ist, erhält man bei einer Vollausteerung im Falle eines rauschartigen Signals

$$\hat{p}_s = \frac{1}{3} \quad , \quad \text{also} \quad (6.23a)$$

$$\text{SNR} = 10 \log \frac{1/3}{2^{-2t}/12} \approx 6,02 t + 6,02 \quad \text{dB} \quad (6.23b)$$

Aus (6.22) ist ersichtlich, daß das SNR eine lineare Funktion der Signalleistung $\hat{p}_s$ ist, solange keine Übersteuerung auftritt und seinen größten Wert gemäß (6.23b) für Vollausteerung annimmt.

Im Falle des Festkommabetragsschneidens läßt sich der absolute Fehler, wie im Unterabschnitt 6.2.2 diskutiert wurde, wie folgt schreiben [68]:

$$\epsilon_b = \epsilon' + \frac{Q_f}{2} \text{sign}[x] \quad , \quad (6.24)$$

wobei ϵ' die gleiche Art statistischer Verteilung wie der absolute Rundungsfehler hat. Wir betrachten ein Sinussignal $\sin(\omega_o t_m + \theta)$. Unter Verwendung der Fourierreihe (6.11) läßt sich (6.24) als

$$\epsilon_b = \frac{2Q_f}{\pi} \sin(\omega_o t_m + \theta) + \epsilon' + \frac{2Q_f}{\pi} \sum_{\mu=1}^{\infty} \frac{\sin[(2\mu+1)(\omega_o t_m + \theta)]}{2\mu+1} \quad (6.25)$$

darstellen. Unter Voraussetzung statistischer Unabhängigkeit von ϵ' und $\sin(\omega_o t_m + \theta)$ läßt sich die Rauschleistung durch

$$\hat{P}_N = Q_f^2 \left[\frac{1}{12} + \frac{1}{4} - \frac{2}{\pi^2} \right] \quad (6.26)$$

ausdrücken [68]. Das SNR berechnet sich dann als

$$SNR_b \approx 6,02 t + 8,84 + 10 \log \hat{p}_s \quad \text{dB} \quad (6.27)$$

Im Falle der Vernachlässigung der statistischen Abhängigkeit zwischen dem absoluten Fehler und dem Signal erhalten wir

$$\hat{P}_N = E[\epsilon_b^2] = 2^{-2t} / 3$$

$$\text{also} \quad SNR_b \approx 6,02 t + 4,77 + 10 \log p_s \quad \text{dB} \quad (6.28)$$

Den beiden Beziehungen (6.27) und (6.28) entnimmt man, daß die Nichtberücksichtigung dieser statistischen Abhängigkeit zu einer zusätzlichen Rauschleistung von etwa 4 dB führt.

Zum sinnvollen Vergleich zwischen Gleit- und Festkomma ist nicht das SNR entscheidend, sondern der Dynamikbereich. Unter Dynamikbereich D versteht man den Abstand (in dB) zwischen dem maximal darstellbaren Signal und dem Rauschen [1,24], also

$$D = 10 \log \frac{\hat{P}_{smax}}{\hat{P}_N} \quad \text{dB} \quad (6.29)$$

wobei P_{smax} die maximale Signalleistung bezeichnet.

Im Falle der Festkommaarithmetik ist jedoch die Leistung des Rauschens vom Wert des Signals nicht abhängig im Gegensatz zur Gleitkommaarithmetik. Um einen fairen Vergleich durchzuführen, betrachten wir das Rauschen bei minimal darstellbarem Signal, also bei niedrigstem Exponenten, im Falle der Gleitkommaarithmetik. Nun betrachten wir den Fall der Festkommaaufrundung. Unter der Annahme, daß das Signal in dem durch (6.20) festgelegten Intervall gleichverteilt ist, läßt sich der Dynamikbereich D_f mit Hilfe von (6.21), (6.23a) und (6.29) ausdrücken als

$$D_f = 10 \log \frac{1/3}{2^{-2t}/12} \approx 6,02 t + 6,02 \quad \text{dB} \quad (6.30)$$

Aus (6.30) und (6.23b) ist ersichtlich, daß D_f und SNR bei einer Vollaussteuerung denselben Wert besitzen.

Im Falle der Gleitkommarundung nimmt man an, daß wiederum das Signal in dem Intervall

$$- 2^{e_{\max}} < x < 2^{e_{\max}}$$

gleichverteilt ist, also

$$\hat{p}_{s\max} = \frac{2^{2e_{\max}}}{12} \quad (6.31)$$

Das Rauschen bei minimalem Exponenten läßt sich mit Hilfe von

$$(\epsilon_{\text{abs}})_{\min} = \epsilon \cdot 2^{e_{\min}}$$

unter Verwendung von (6.5) und (6.7) schreiben als

$$\hat{p}_N = \sigma_c^2 \cdot 2^{2e_{\min}} = \frac{2^{-2t_{\max}}}{12} \cdot 2^{2e_{\min}} \quad (6.32)$$

Setzt man (6.31) und (6.32) in (6.29) ein, so läßt sich der Dynamikbereich D_g ausdrücken durch

$$D_g \approx 6,02 (e_{\max} - e_{\min} + t_{\max}) \text{ dB} \quad (6.33)$$

Betrachtet man den Fall, wenn der Exponent in Vorzeichenbetrag dargestellt ist, also $e_{\min} = -e_{\max}$, so erhält man

$$D_g \approx 6,02 (2^{t_{\exp} + 1} + t_{\max} - 2) \text{ dB} \quad (6.34)$$

Aus (6.34) und (3.34) ist zu ersehen, daß $D_g = ZB_{g2}$ ist.

Es sei bemerkt, daß der Dynamikbereich bei Gleitkomma mit der Wortlänge des Exponenten exponentiell zunimmt. Das bedeutet, daß durch geeignete Wahl der Exponentenwortlänge der Dynamikbereich weitaus besser als bei Festkomma. Als Beispiel nehme man an, daß $t=15$, $t_{\exp}=4$ und $t_{\max}=10$ Bits jeweils ohne Vorzeichenbit. Wir erhalten $D_f \approx 96 \text{ dB}$ und $D_g \approx 241 \text{ dB}$.

6.4 Berechnung des Rundungsrauschens eines beliebigen Wellendigitalfilters

Wellendigitalfilter verhalten sich im Hinblick auf das Rundungsrauschen prinzipiell wie der zuvor untersuchte Quantisierer (mit der Ausnahme, daß das SNR im Falle eines Filters jedoch in der Regel ungünstiger als das SNR eines Quantisierers ist). Unter der Voraussetzung unkorrelierter Rauschquellen läßt sich das SNR am Ausgang eines Digitalfilters in geschlossener Form ermitteln [35,70,71]. Im folgenden betrachten wir ein WDF mit Gleitkommaarithmetik und verwenden den absoluten Fehler. Das gesamte Rauschen am Ausgang des Filters wird aus der überlagerten Wirkung aller Rauschquellen auf das System berechnet.

In Kapitel 4 haben wir gesehen, daß die durch Begrenzung der Signalwortlänge in einem WDF hervorgerufenen Fehlerquellen an allen Adaptorausgängen einwirken. Man kann demnach ein nichtlineares WDF durch ein lineares System mit zusätzlichen Fehlerquellen ϵ_{absu} ($u=1,2,3,\dots,n$) gemäß Bild 6.3 beschreiben.

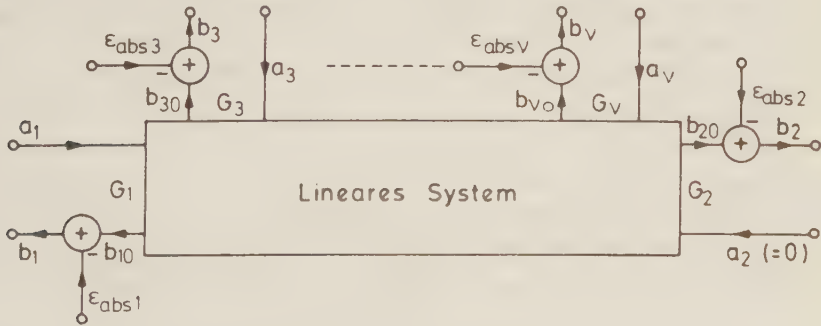


Bild 6.3: Ein WDF mit Rauschquellen ϵ_{absu} .

Sei $h_{2u}(m)$ die Impulsantwort und $H_{2u}(z)$ die Übertragungsfunktion vom Knoten u zum Filterausgang (z.B. b_2), dann gilt für die Varianz des Gesamtfehlers

$$\sigma_N^2 = \sum_{u=2}^n \sigma^2 \epsilon_{absu} \sum_{m=0}^{\infty} h_{2u}^2(m) \quad , \quad (6.35)$$

unter Voraussetzung unkorrelierter Rauschquellen mit $E[\epsilon_{absu}] = 0$, $E[\epsilon_{absu}(mT) \epsilon_{absu}(mT + lT)] = 0$ für $l \neq 0$ und $E[\epsilon_{absu}^2] = \sigma^2$.

Mit den Regeln der Z-Transformation erhalten wir die bekannte Formel

$$\sum_{m=0}^{\infty} h_{2v}^2(m) = \frac{1}{2\pi j} \oint_{|z|=1} H_{2v}(z) H_{2v}(1/z) z^{-1} dz = \|H_{2v}\|_2^2, \quad (6.36)$$

wobei $\|H_{2v}\|_2$ die L_2 -Norm ist. Durch Einsetzen von (6.36) in (6.35) erhalten wir

$$\sigma_N^2 = \sum_{v=2}^n \sigma_{\epsilon_{absv}}^2 \|H_{2v}\|_2^2. \quad (6.37)$$

Als nächstes betrachten wir die in (6.37) auftretenden Varianzen $\sigma_{\epsilon_{absv}}^2$ der einzelnen Rauschquellen. Es gilt also für die absoluten Fehler die folgende Bezeichnung:

$$\epsilon_{absv} = \epsilon_{rv} b_{v0} \quad (6.38)$$

Mit der bereits im unter Abschnitt 6.2.1 erwähnten Voraussetzung, daß die statistische Abhängigkeit von ϵ_{rv} und b_{v0} vernachlässigbar ist, erhalten wir

$$\sigma^2 = \sigma_{\epsilon_{rv}}^2 \cdot \sigma_{b_{v0}}^2 = \sigma_{\epsilon_r}^2 \cdot \sigma_{b_{v0}}^2 \quad (6.39)$$

Wir setzen im folgenden voraus, daß am Eingang des Filters ein weißes Rauschsignal mit $E[a_1]=0$ und $E[a_1^2]=\sigma_x^2$ liegt.

Sei $h_{v1}(m)$ die Impulsantwort und $H_{v1}(z)$ die Übertragungsfunktion vom Filtereingang zum Knoten v , dann gilt

$$\sigma_{b_{v0}}^2 = \sigma_x^2 \cdot \|H_{v1}\|_2^2, \quad (6.40a)$$

$$\sigma_{b_2}^2 = \sigma_x^2 \cdot \|H_{21}\|_2^2. \quad (6.40b)$$

Unter Verwendung von (6.40) und (6.39) läßt sich die Varianz des Gesamtfehlers (Gl.(6.37)) als

$$\sigma_N^2 = \sigma_{\epsilon_r}^2 \cdot \sigma_{b_2}^2 \cdot K_f \quad (6.41)$$

darstellen, wobei

$$K_f = \sum_{v=2}^n \frac{||H_{v1}||_2^2 \cdot ||H_{2v}||_2^2}{||H_{21}||_2^2} \quad (6.42)$$

eine Konstante ist, welche nur von der Filterschaltung abhängt. Das SNR am Filterausgang erhält unter Verwendung der Tabelle 3.3 die folgende Gestalt

$$SNR_g \cong 6,02 t_{man} + 7,78 - 10 \log K_f \quad (6.43)$$

Im Falle der Festkommaarithmetik läßt sich das SNR als

$$SNR_f \cong 6,02t + 10,79 + 10 \log \hat{P}_s - 10 \log \left[\sum_{v=2}^n ||H_{2v}||_2^2 \right] \quad (6.44)$$

ausdrücken.

In der vorhergehenden Diskussion haben wir gesehen, daß das SNR am Filterausgang durch (6.43) und (6.44) ermittelt werden kann. Allerdings erfolgt die Berechnung der hierzu erforderlichen L_2 -Normen üblicherweise durch Ermittlung und Integration der zugehörigen Übertragungsfunktionen (vgl. 6.36). Dies kann aber auch, wie wir in (6.36) gesehen haben, über die entsprechenden Impulsantworten berechnet werden. Es ist jedoch nicht notwendig, diese Impulsantworten explizit zu berechnen, da sich die gesuchten L_2 -Normen durch eine Matrixbeziehung in Abhängigkeit von den Systemparametern ausdrücken lassen [20].

Anhand eines Beispiels werden wir das SNR eines WDF's durch Simulation in Gleitkommaarithmetik untersuchen und mit dem SNR bei Festkommaarithmetik vergleichen.

6.5 Beispiel

Wir betrachten einen Tiefpaß 3.Ordnung (Bild 4.19). Dieses Filter wurde auf dem Rechner HP1000 in Gleitkommaarithmetik simuliert. Die Signalparameter wurden aus dem Gleitkommasystem GS(2,10,-16,15) gewählt. Es wurde am Filtereingang ein Sinussignal verwendet (Frequ.=5 KHz.). Außerdem wurde Rundung an allen Adaptorausgängen angewandt. Parallel hierzu wurde eine identische Filter-

schaltung mit gleichem Eingangssignal aber mit hinreichender Wortlänge simuliert (25 Bit für die Mantissa und 7 Bit für den Exponenten jeweils ohne Vorzeichenbit). Das Fehlersignal wurde entsprechend Bild 6.4 bestimmt, wobei H_0 das nichtlineare Filter und H_∞ das mit hinreichender Wortlänge simulierte Filter bezeichnet.

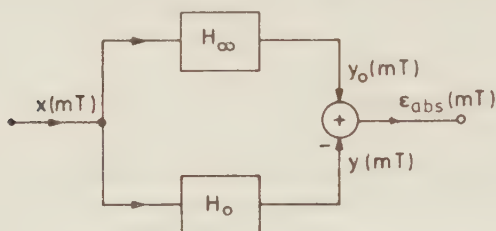


Bild 6.4: Bestimmung des Rundungsrauschens.

Entsprechend Bild 6.4 sei $y(mT)$ die Ausgangsfolge des "realen" Filters und $y_0(mT)$ die Ausgangsfolge des "idealen" Filters. Gemäß (3.22) läßt sich die Fehlerfolge als

$$\epsilon_{abs}(mT) = y_0(mT) - y(mT) \quad (6.45)$$

definieren. Die zugehörige Rauschleistung für einen Zeitraum von $mT=0$ bis $mT=MT$ ($M=3000$) wurde durch

$$\hat{P}_N = \frac{1}{(M+1)T} \sum_{m=0}^M \epsilon_{abs}^2(mT) \quad (6.46)$$

bestimmt. Auf ähnliche Weise wurde die Signalleistung durch

$$\hat{P}_S = \frac{1}{(M+1)T} \sum_{m=0}^M y_0^2(mT) \quad (6.47)$$

berechnet. Mit Hilfe von (6.47) und (6.46) wurde das SNR ermittelt.

Um einen sinnvollen Vergleich zwischen Gleit- und Festkommaarithmetik bezüglich des SNR's durchzuführen, muß der Hardwarerealisierungsaufwand des Filters (hier des 2-Tor-Adaptors) in beiden Fällen derselbe bleiben. Für einen 2-Tor-Adaptor werden im Falle

der Gleitkommaarithmetik 4 Multiplikationen und 2 Additionen und im Falle der Festkommaarithmetik 2 Multiplikationen und 3 Additionen benötigt. Unter Verwendung von (3.37), (3.38), (3.39) und (3.40) wurde die Signalwortlänge für Festkommarealisierung mit 16 Bit (ohne Vorzeichenbit) gewählt.

Es wurde dann Rundung an allen Adaptoraustritten angewandt. Das Rauschen wurde entsprechend Bild 6.4 bestimmt. Bild 6.5 zeigt den Verlauf des SNR's über der Signalleistung A.

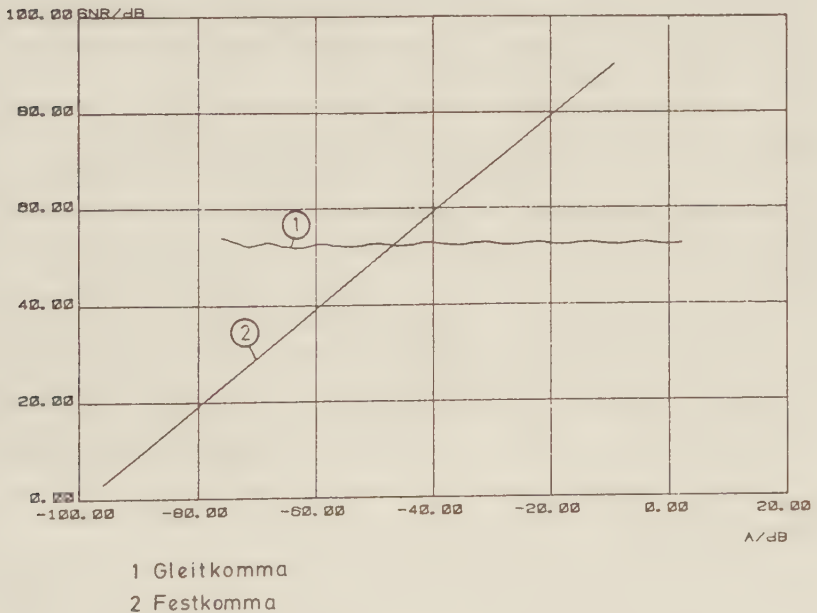


Bild 6.5: SNR eines Tiefpasses 3.Ordnung.

Wie aus Bild 6.5 zu ersehen ist, hat die Gleitkommaarithmetik bei kleineren Signalen ein überlegenes SNR Verhalten gegenüber der Festkommaarithmetik. Will man aber auf den Realisierungsaufwand verzichten und ein günstiges SNR für alle Signalamplituden erreichen, so muß man die Mantissenwortlänge dementsprechend groß wählen.

7. Zusammenfassung

Das wesentliche Ziel dieser Arbeit ist es, Wellendigitalfilter mit Gleitkommaarithmetik zu realisieren, d.h., daß sowohl Multipliziererkoeffizienten als auch Signalparameter als Gleitkommagrößen dargestellt werden. Wir beschränken uns auf den praktischen interessanten Fall, nämlich den Fall der Darstellung einer Zahl als Gleitkommagröße in Zweierkomplementkode.

Im 2. Kapitel wurden zunächst die wichtigsten für das Verständnis dieser Arbeit erforderlichen Grundlagen aus der Theorie der Wellendigitalfilter sowie einige günstige Eigenschaften dieser Filter rekapituliert. Hierzu zählen insbesondere die geringe Empfindlichkeit der Dämpfung gegenüber Koeffizientenänderungen und das Stabilitätsverhalten solcher Filter.

Im 3. Kapitel wurden die Gleitkommadarstellung sowie die Quantisierungsfehler und der Zahlenbereich näher betrachtet.

Im 4. Kapitel wurden mit Hilfe des Konzepts der Pseudopassivität Maßnahmen angegeben, die bewirken, daß nach Abschalten des Eingangssignals alle auftretenden parasitären Schwingungen in einem Gleitkomma-Wellendigitalfilter abklingen, sogar wenn das Filter in einer Nachrichtenschleife arbeitet. Außerdem wurde mit dem Konzept der inkrementalen Pseudopassivität nachgewiesen, daß bei Verwendung geeigneter Überlaufcharakteristiken (unter Vernachlässigung der Signalgranularität) Gleitkomma-Wellendigitalfilter auch antwortstabil sind. Es wurden Simulationen zur Demonstration der hervorragenden Stabilitätseigenschaften von Wellendigitalfiltern unter unterschiedlichen Betriebsbedingungen durchgeführt.

Im 5. Kapitel wurden Stabilitätsuntersuchungen von konventionellen rekursiven Digitalfiltern 2. Ordnung durchgeführt. Diese Untersuchungen orientierten sich im wesentlichen an der Stabilität sowohl bei abgeschaltetem Eingangssignal als auch unter Schleifenbedingungen. Es hat sich herausgestellt, daß in solchen Filtern unter den eben genannten Betriebsbedingungen Grenzyklusschwingungen mit teilweise großen Amplituden entstehen können. Allerdings würde die Situation noch weitaus kritischer, wenn das Filter aus mehreren Gliedern 2. Ordnung besteht, wie dies in der Praxis der Fall ist. Außerdem wurde in diesem Kapitel ein auf Wellendigitalfilter beruhendes Gleitkommadigitalfilter 2. Ordnung besprochen, welches antwortstabil ist und bei abgeschaltetem Eingangssignal frei von Grenzyklusschwingungen.

Im 6. Kapitel wurde das Rauschen am Ausgang eines Digitalfilters mit Hilfe des Modells unkorrelierter Rauschquellen untersucht. Es hat sich herausgestellt, daß das Signal-Rausch-Verhältnis bei Gleitkommaarithmetik ein vorteilhaftes Verhalten gegenüber Festkommaarithmetik besitzt, wenn der Aufwand der Hardwarerealisierung in Kauf genommen wird. Darüber hinaus ist der Dynamikbereich bei Gleitkomma weitaus besser als bei Festkomma.

Ich danke Herrn Prof. Dr. A. Fettweis für die interessante Aufgabenstellung und für wertvolle Anregungen.

Mein Dank gilt weiterhin Herrn Prof. Dr. K. Meerkötter für die Übernahme des Korreferats.

Literaturverzeichnis

- [1] A. Fettweis, "Wave digital filters: theory and practice, Proc. IEEE, Bd. 47, Nr. 2, S. 270-327, Feb. 1986.
- [2] A. Fettweis, "Digital filter structures related to classical filter networks", Archiv für Elektronik und Übertragungstechnik, Bd. 25, S. 79-89, Feb. 1971.
- [3] A. Fettweis, "Some principles of designing digital filters imitating classical filter structures", IEEE Trans. Circuit Theory, CT-18, S. 314-316, März 1971.
- [4] A. Fettweis, "Entwurf von Digitalfiltern in Anlehnung an Verfahren der klassischen Netzwerktheorie", Tagungsheft der NTG-Fachtagung "Analyse und Synthese von Netzwerken", Stuttgart, S. 17-20, Okt. 1970.
- [5] A. Fettweis, "Reciprocity, interreciprocity, and transposition in wave digital filters", Int. Jour. of Circuit Theory and Application, Bd. 1, Nr. 4, S. 323-337, Okt. 1973.
- [6] A. Sedlmeyer und A. Fettweis, "Digital filters with true ladder configuration", Int. Jour. of Circuit Theory and Applications, Bd. 1, Nr. 1, S. 5-10, Januar 1973.
- [7] A. Fettweis, "Realizability of digital filter networks", Archiv für Elektronik und Übertragungstechnik, Bd. 30, S. 90-96, Feb. 1976.
- [8] A. Fettweis, "Filters met willekeurig gekozen dempingspolen en Tschebyschewkarakteristiek in het doorlaatgebied, "Tijdschrift van het Nederlands Radiogenootschap, Bd. 25, Nr. 5-6, S. 337-382, 1960.
- [9] H.J. Orchard, G.C. Temes und T. Cataltepe, "General sensitivity theorems for terminated lossless two-ports", Proc. 1984, IEEE Int. Symp. Circuits and Systems, S. 177-180, 1984.
- [10] H.J. Orchard, G.C. Temes und T. Cataltepe, "Sensitivity formulas for terminated lossless two-ports", IEEE Trans. Circuits and Systems, Bd. CAS-32, Nr. 5, S. 459-466, Mai 1985.
- [11] A. Fettweis, Vorlesungsskript über "Grundlagen der Theorie elektrischer Schaltungen", Ruhr-Universität Bochum.
- [12] A. Fettweis, "On the connection between multiplier word length limitation and roundoff noise in digital filters", IEEE Trans. Circuit Theory, Bd. CT-19, S. 486-491, Sept. 1972.
- [13] A. Fettweis, "Realisierungsprobleme von Digitalfiltern - Durch Wortlängenbegrenzung bedingte Effekte sowie Filterstrukturen", NTZ-Archiv, Bd. 2, S. 229-233, 1980.

- [14] K.A. Owenier, "Fast computation of sensitivities in wave digital filters", ECCTD-81, S. 718-722, 1981.
- [15] K.A. Owenier, "A general method for efficient computation of sensitivities in wave digital filters", IEEE Trans. Circuits and Systems, Bd. CAS-30, s. 750-757, Okt. 1983.
- [16] A. Fettweis, G.J. Mandiville und C.Y. Kao, "Design of wave digital filters for communication applications", Proc. 1975 IEEE Int. Symp. Circuits and Systems, S. 162-165, Newton, Mass., Apr. 1975.
- [17] K.A. Owenier, "Optimierung von Filtern, insbesondere von Wellendigitalfiltern mit verringerter Zahl an Multiplizierern", Dissertation, Ruhr-Universität Bochum, 1977.
- [18] W. Wegener, "Design of wave digital filters with very short coefficient word lengths", Proc. 1976 IEEE Int. Symp. Circuits and Systems, S. 473-476, München, Apr. 1976.
- [19] W. Wegener, "On the design of wave digital lattice filters with short coefficient word lengths and optimal dynamic range", IEEE Trans. Circuits and Systems, Bd. CAS-25, S. 1091-1098, Dez. 1978.
- [20] W. Wegener, "Entwurf von Wellendigitalfiltern mit minimalem Realisierungsaufwand", Dissertation, Ruhr-Universität Bochum, 1980.
- [21] A. Fettweis, "Roundoff noise and attenuation sensitivity in digital filters with fixed-point arithmetic", IEEE Trans. Circuit Theory, Bd. CT-20, Nr. 2, S. 174-175, März 1973.
- [22] A. Fettweis, "On sensitivity and roundoff noise in wave digital filters", IEEE Trans. Acoust. Speech and Signal Processing, Bd. ASSP-22, Nr. 5, S. 383-384, Okt. 1974.
- [23] K. Renner und S.C. Gupta, "Reduction of roundoff noise in wave digital filters", IEEE Trans. Circuits and Systems, Bd. CAS-21, Nr. 2, S. 305-310, März 1974.
- [24] A. Fettweis, "Digital circuits and systems", IEEE Trans. Circuits and Systems, Bd. CAS-31, Nr. 1, S. 31-48, Jan. 1984.
- [25] A. Fettweis, "Pseudopassivity, sensitivity, and stability of wave digital filters", IEEE Trans. Circuit Theory, Bd. CT-19, Nr. 6, S. 668-673, Nov. 1972.
- [26] A. Fettweis und K. Meerkötter, "Suppression of parasitic oscillations in wave digital filters", IEEE Trans. Circuits and Systems, Bd. CAS-22, Nr. 3, S. 239-246, März 1975.
- [27] A. Fettweis und K. Meerkötter, "Correction to "Suppression of parasitic oscillations in wave digital filters", IEEE Trans. Circuits and Systems, Bd. CAS-22, S. 575, Juni 1975.

- [28] T.A.C.M. Claasen, W.F.G. Mecklenbräuker und J.B.H. Peek, "On the stability of the forced response of digital filters with overflow nonlinearities", IEEE Trans. Circuits and Systems, Bd. CAS-22, Nr. 8, S. 692-696, Aug. 1975.
- [29] K. Meerkötter, Beiträge zur Theorie der Wellendigitalfilter, Dissertation, Ruhr-Universität Bochum, 1980.
- [30] K. Meerkötter, "Incremental pseudopassivity of wave digital filters", Proc. First European Signal Processing Conference (EUSIPCO-80), S. 27-32, Lausanne, Switzerland, Sept. 1980.
- [31] U. Kulisch, "Grundlagen des Numerischen Rechners, Mathematische Begründung der Rechnerarithmetik", Wissenschaftsverlag, Bibliographisches Institut, 1976.
- [32] U. Kulisch und W.L. Miranker, "Computer Arithmetik in theory and practice", Academic Press, 1981.
- [33] A. Fettweis, "Some statistical properties of floating-point roundoff noise", Proc. Sixteenth Midwest Symp. Circuit Theory, Waterloo, Ontario, Canada, Apr. 12-13, 1973, S. II.2.1-II.2.10.
- [34] A. Fettweis, "On properties of floating-point roundoff noise", IEEE Trans. Acoust., Speech, and Signal Processing, Bd. ASSP-22, Nr. 2, S. 149-151, Apr. 1974.
- [35] B. Liu und T. Kaneko, "Error analysis of digital filters realized with floating-point arithmetic", Proc. IEEE, Bd. 57, S. 1735-1747, Okt. 1969.
- [36] J.A. Eldon und C. Robertson, "A floating-point format for signal processing", Proc. IEEE Int. Conf. on Acoust., Speech, and Signal Processing, Paris 1982, S. 717-720.
- [37] B. Woo, L. Liu und F. Ware, "A high-speed 32 bit IEEE floating point chip set for digital signal processing, Proc. IEEE ISCAS 1984, S. 16.6.1-16.6.4.
- [38] D. Mitra und J.R. Boddie, "Limit cycles in floating point digital filters", Proc. ISCAS 1979, S. 374-377.
- [39] Produktbeschreibung von NEC PD 77230.
- [40] A. Lacroix, "Digitale Filter", Oldenburg-Publisher, München-Wien, 1980.
- [41] A. Lacroix, und R. Block, "Floating-point computation in digital signal processing, Proc. IEEE ISCAS 1983, S. 709-712.
- [42] S.R. Parker und S.F. Hess, "Limit cycle oscillations in digital filters", IEEE Trans. Circuit Theory, Bd. CT-18, S. 687-697, Nov. 1971.

- [43] A.N. Willson, JR, "Limit cycles due to adder overflow in digital filters", IEEE Trans. Circuit Theory, Bd. CT-19, S. 342-346, Juli 1972.
- [44] J.F. Kaiser, "On the limit cycle problem", IEEE Int. Symp. Circuits and Systems, München, S. 642-645, Apr. 1976.
- [45] T.A.C.M. Claasen, W.F.G. Mecklenbräuker und J.B.H. Peek, "Effects of quantization and overflow in recursive digital filters", IEEE Trans. Acoust., Speech, and Signal Processing, Bd. ASSP-24, S. 517-529, Dez. 1976.
- [46] P.M. Ebert, J.E. Mazo und M.G. Taylor, "Overflow oscillations in digital filters", Bell Syst. Tech. J., Bd. 48, S. 2999-3020, Nov. 1969.
- [47] T.Kaneko, "Limit cycle oscillations in floating point digital filters", IEEE Trans. Audio & Electroacoustics, Bd. AU-21, Nr. 2, S. 100-106, Apr. 1973.
- [48] A. Lacroix, "Limit cycles in floating point digital", AEÜ, Bd. 30, Heft 718, S. 277-284, Aug. 1976.
- [49] A. Fettweis und K. Meerkötter, "Suppression of parasitic oscillations in half-synchronous wave digital filters", IEEE Trans. Circuits and Systems, Bd. CAS-23, S. 125-126, Febr. 1976.
- [50] A. Fettweis und K. Meerkötter, "On parasitic oscillations in digital filters under looped conditions", IEEE Trans. Circuits and Systems, Bd. CAS-24, S. 475-481, Sept. 1977.
- [51] A. Fettweis und K. Meerkötter, "On adaptors for wave digital filters", IEEE Trans. Acoust., Speech, and Signal Processing, Bd. ASSP-23, Nr. 6, S. 516-525, Dez. 1975.
- [52] C.D.R. de Vaal und R. Nouta, "On the suppression of parasitic oscillations in floating point wave digital filters", IEEE Trans. Circuits and Systems, Bd. CAS-27, Nr. 2, S. 144-145, Febr. 1980.
- [53] G. Bohlender und K. Grüner, "Gesichtspunkte zur Implementierung einer optimalen Arithmetik, Berichte des German chapter of the ACM 10, (Hrsg.) U. Kulisch/Ch. Ullrich.
- [54] A. Fettweis, H. Levin und A. Sedlmeyer, "Wave digital lattice filters", Int. Journal of Circuit Theory and Applications, Bd. 2, Nr. 2, S. 203-211, Juni 1974.
- [55] R. Nouta, "The Jaumann structure in wave digital filters", Int. Journal of Circuit Theory and Applications, Bd. 2, Nr. 2, S. 163-174, Juni 1974.
- [56] W. Wegener, "Wave digital directional filters with reduced number of multipliers and adders", Archiv für Elektronik und Übertragungstechnik, Bd. 33, Nr. 6, S. 239-243, Juni 1979.

- [57] A. Fettweis, J.A. Nossek und K. Meerkötter, "Reconstruction of signals after filtering and sampling rate reduction", IEEE Trans. Acoust., Speech, and Signal Processing, Bd. ASSP, Nr. 4, S. 893-902, Aug. 1985.
- [58] L. Gazsi, "Explicit formulas for lattice wave digital filters", IEEE Trans. Circuits and Systems, Bd. CAS-32, S. 68-88, Jan. 1985.
- [59] V. Belevitch, "Classical Network Theory", San Francisco, CA: Holden Day, 1968.
- [60] W. Drews, "Entwurf und Aufbau digitaler Polyphasen-Filterbänke unter Verwendung von Wellendigitalfiltern", Dissertation, Ruhr-Universität Bochum, 1986.
- [61] S.N. Güllüoglu, "Diskrete Optimierung von Wellendigitalfiltern, insbesondere von Brücken-Wellendigitalfiltern mit Zweitoradaptoren", Dissertation, Ruhr-Universität Bochum, 1986.
- [62] R. Nouta, "Wave digital cascade synthesis", Int. Journal of Circuit Theory and Applications, Bd. 3, S. 231-247, Juli 1975.
- [63] D. Mitra, "A bound on limit cycles in digital filters which exploits a particular structural property of the quantization", IEEE Trans. Circuits and Systems, Bd. CAS-24, Nr. 11, Nov. 1977.
- [64] K. Meerkötter, W. Wegener, "A new second-order digital filter without parasitic oscillations", AEÜ, Bd. 29, S. 312-314, 1975.
- [65] C. Verkroost, H.J. Butterweck, "Suppression of parasitic oscillations in wave digital filters and related structures by means of controlled rounding", AEÜ, Bd. 30, S. 181-186, 1976.
- [66] A. Cervera, "Untersuchung der Stabilität konventioneller rekursiver Digitalfilter unter Schleifenbedingungen", Dissertation, Ruhr-Universität Bochum, 1987.
- [67] A. Fettweis, "On the evaluation of roundoff noise in digital filters", IEEE Trans. Circuit and Systems, Bd. CAS-22, S. 896, Nov. 1975.
- [68] K. Meerkötter, "Aufbau und Untersuchungen eines Wellendigitalfilters mit echter Abzweigstruktur", Diplomarbeit, Lehrstuhl für Nachrichtentechnik, Ruhr-Universität Bochum, 1973.
- [69] U. Ullrich, "Effektive Verfahren zur Ermittlung des Signalrauschverhältnisses von Digitalfiltern unterschiedlicher Struktur", Dissertation, Ruhr-Universität Bochum, 1976.
- [70] C. Weinstein, A.V. Oppenheim, "A comparison of roundoff noise in floating-point and fixed-point digital filters realizations", Proc. IEEE, Bd. 57, S. 1181-1183, Juni 1969.

- [71] E.P.F. Kan, J.K. Aggarwal, "Error analysis of digital filters employing floating-point arithmetic", IEEE Trans. Circuit Theory, Bd. CT-18, Nr. 6, S. 678-686, Nov. 1971.
- [72] W.H. Ku, S.M. Ng, "Floating-point coefficient sensitivity and roundoff noise of recursive digital filters realized in ladder structures", IEEE Trans. Circuit Theory and Systems, Bd. CAS-22, Nr. 12, S. 927-936, Dez. 1975.

Lebenslauf

17.8.1949	geboren in Menofia/Ägypten
1954-1960	Besuch der Grundschule in Mescherif/Ägypten
1960-1964	Besuch der Vorbereitungsschule auf das Gymnasium in Kairo
1964-1969	Besuch des Gymnasiums in Kairo
Juni 1969	Abitur
1969-1974	Studium der Elektrotechnik an der Al-Azhar Universität in Kairo
Juni 1974	B.Sc. in Elektrotechnik
31.12.1977	M.Sc. in Elektrotechnik
1974-1982	Wissenschaftlicher Assistent an der Abtei- lung für Elektrotechnik der Al-Azhar Univer- sität Kairo
April 82-1.10.1982	Sprachkurs im Goethe-Institut in Lüneburg
seit Oktober 1982	Doktorand am Lehrstuhl für Nachrichtentech- nik der Ruhr-Universität Bochum

